Logic and Interactive RAtionality

Yearbook 2010

The printing of this book was supported by the UP fund of Johan van Benthem.

Cover design by Nina Gierasimczuk.

Foreword

This yearbook is the third in the series of the LIRa yearbooks started in 2008 (when the seminar was still known as "Logics for Dynamics of Information and Preferences"). Three years might not yet be called a "tradition", but 2010 has definitely witnessed the "beginning of a beautiful friendship" within the seminar, with sessions regularly held between Amsterdam (hosted by the Institute for Logic, Language and Computation of the University of Amsterdam) and Groningen (hosted by the Department of Theoretical Philosophy of the University of Groningen). This has turned a local initiative into a truly national event.

The volume contains papers on a rich array of topics directly reflecting the seminar's livelihood in 2010. We greatly thank the authors for their excellent contributions and for their cooperation during the whole editing process of the yearbook. Many thanks go to Johan van Benthem for stimulating and supporting, both intellectually and financially, the seminar and this yearbook. Finally, we want to thank Nina Gierasimczuk for yet another captivating cover, and Alexandru Marcoci and Fernando Velázquez-Quesada for their precious LaTeX help.

Amsterdam
June 11, 2011

Davide Grossi
Ştefan Minică
Ben Rodenhäuser
Sonja Smets
(eds.)

Content

Preface

Johan van Benthem

Many things may be said about the passing years, but one good thing is that they produce Yearbooks. The volume that you are about to enter is a lively collection that defies easy description, though one clear trend appears to be dynamics fanning out into new territories. There are pieces on the logic of probabilistic agreement scenarios, counterfactual reasoning, epistemic closure in epistemology, protocol analysis, strategies in game theory, and a pleasant surprise: medieval obligatio games. And to that already striking range, the book adds further pieces with suggestive new perspectives on strength of beliefs, design and future knowledge, as well as binary aggregation problems in social choice theory. And the fare is even richer than this menu, once you realize how several former Yearbook contributors have defended dissertations in 2010, on dynamic logics of belief change, game theory, and learning theory, which you can consult at the seminar's website `www.illc.uva.nl/lgc/seminar/`. This document invites you to learn about this world, but more than that, it also invites you to participate.

Johan van Benthem
June 2011

Agreement Theorems in Probabilistic Dynamic Epistemic Logic

Lorenz Demey

University of Leuven
lorenz.demey@hiw.kuleuven.be

Abstract

This paper studies Aumann's agreeing to disagree theorem from the perspective of probabilistic dynamic epistemic logic. We introduce enriched probabilistic Kripke frames and models, and various ways of updating them. This framework is then used to prove several agreement theorems, which are natural formalizations of Aumann's original result. Furthermore, we provide a sound and complete axiomatization of a dynamic agreement logic, in which one of these agreement theorems can be derived syntactically. These technical results are then used to clarify some conceptual issues surrounding the agreement theorem, in particular the role of common knowledge and the importance of explicitly representing the dynamics.

1 Introduction

The main goal of this paper is to study Aumann's celebrated 'agreeing to disagree' theorem (Aumann 1976) from the perspective of epistemic logic, in particular *probabilistic dynamic epistemic logic* (PDEL). The agreement theorem, and the related no-trade theorem (Milgrom and Stokey 1982) are of central importance in game theory. Several notions connected to this theorem, such as the common prior assumption, and, especially, the notion of common knowledge,

have been studied extensively by game theorists, but also by philosophers, computer scientists and logicians (Halpern and Moses 1990, Lewis 1969, Milgrom and Stokey 1982). This paper thus establishes a new connection between the epistemic-logical and game-theoretical perspectives on (common) knowledge and related epistemic notions.

This endeavor also has definite advantages for both epistemic logic and game theory as separate disciplines. Probabilistic DEL is a recent development, and to capture the agreement theorems in this framework, several extensions and improvements were necessary. For example, we introduce a new way of defining updated probability functions, which elegantly avoids several of the problems mentioned by Kooi (2003), and is thus also of independent interest. The logical perspective on the agreement theorems has definite advantages for game theorists as well. The technical results established in this paper lead to philosophical and methodological clarifications of some issues surrounding the agreement result. In particular, it will be argued that the role of common knowledge is less central than is often thought, and that explicitly representing the dynamics is essential to obtain the most natural agreement theorems.[1]

The remainder of this paper is organized as follows. Section 2 provides an introduction to Aumann's original agreement theorem and highlights those features that will become particularly important in later sections. In Section 3 we introduce the semantic setup of probabilistic dynamic epistemic logic. We define (enriched) probabilistic Kripke frames and models, and we introduce three ways of updating them: (1) carrying out experiments, (2) public announcement of a formula φ, and (3) a *dialogue* about a formula φ, i.e. a sequence of public annoucements that reaches a fixed point after finitely many steps. Section 4 contains the key results of this thesis, viz. several (dynamic) agreement theorems for probabilistic Kripke models/frames. Section 5 provides characterization results for all conditions of the agreement theorems, and then uses these to obtain a sound and complete agreement logic. Section 6 uses the formal results to shed some new light on two important philosophical/methodological issues: we will argue that common knowledge is (at least conceptually speaking) not so central for agreeing to disagree results as is often thought, and secondly, our results seem to show that the only 'natural' agreement theorems are all dynamic

[1] Recently, Dégremont and Roy (2009) have brought Aumann's agreement theorem (and some extensions) already to epistemic logic. They, however, didn't use probabilistic Kripke models, but rather (qualitative) epistemic plausibility models. A detailed comparison between Dégremont and Roy's approach and our approach is outside the scope of this paper, but can be found in Demey (2010). There it is argued that our probabilistic approach is to be preferred on both philosophical and technical grounds.

in nature, and that static theorems (such as Aumann's original one) are only possible at the expense of a convoluted semantic setup. Section 7 wraps things up.

2 Aumann's original agreement theorem

Aumann originally expressed the 'agreeing to disagree' theorem as follows: "If two people have the same prior, and their posteriors for an event A are common knowledge, then these posteriors are equal." (Aumann 1976, p. 1236). In other words: if two people have the same prior, then they cannot *agree* (have common knowledge of their posteriors) *to disagree* (while these posteriors are not equal). It is clear that, when phrased in this way, the agreement theorem is a *static* result: it is a conditional statement that can be expressed without any dynamic operators: $[\mathsf{equalpriors} \wedge C(\mathsf{posteriors})] \rightarrow \mathsf{equalposteriors}$.

Aumann also motivates his theorem by sketching an informal scenario that embodies the intuitions behind it.[2] Roughly speaking the scenario looks as follows. We are considering two agents, 1 and 2. Initially, they have the same probability distribution ($P_1 = P_2$). Then both agents separate and each perform an experiment. Immediately afterwards, the agents' probability distributions have changed due to the information that they have gained from their experiments. Because the agents performed different experiments, their probability distributions have changed in different ways. In particular, for some φ, it holds that $P_1(\varphi) = a$ and $P_2(\varphi) = b$ (for some $a, b \in [0, 1]$), while $a \neq b$. Furthermore, since agent 1 doesn't know the outcome of agent 2's experiment, she doesn't know how agent 2's probability function has changed. A symmetric argument applies to agent 2. Hence at this stage it is not common knowledge between both agents that $P_1(\varphi) = a$ and $P_2(\varphi) = b$. Finally, the agents start communicating with each other. Agent 1 tells agent 2 that $P_1(\varphi) = a$; on the basis of this new information, agent 2 changes her probability function, which she, in turn, communicates to agent 1, etc. At a certain point in the conversation, the agents obtain common knowledge of their probabilities. Since both agents had the same prior ($P_1 = P_2$ initially) and their posteriors have become common knowledge, Aumann's theorem now says that these probabilities have to coincide.

Although the formal agreement theorem is a static result, the intuitive scenario behind it clearly involves several dynamic phenomena. Two broad types

[2] A similar explanatory scenario is described more extensively by Bonanno and Nehring (1997).

of dynamics can be distinguished: (1) the *experiments* and (2) the *communication*.

3 The General Setup of PDEL

In this section we introduce the general semantic setup of probabilistic dynamic epistemic logic. This setup will be used in Section 4 to formalize and prove various dynamic agreement theorems.

3.1 Probabilistic Kripke models

We first introduce (enriched) probabilistic Kripke frames and models. We focus on the two agent-case (this will suffice for the statement of the agreement theorems); generalizations to any (finite) number of agents are straightforward. We also fix a countably infinite set *Prop* of proposition letters.

Definition 3.1. An *(enriched) probabilistic Kripke frame* (for two agents) is a tuple $\mathbb{F} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2 \rangle$, where W is a non-empty finite set of states, R_1, R_2, E_1 and E_2 are equivalence relations, and μ_1 and μ_2 assign to each world $w \in W$ a probability mass function $\mu_i(w) : W \to [0, 1]$ (so $\sum_{v \in W} \mu_i(w)(v) = 1$). We also require (i) that $\mu_i(w)(w) > 0$ for all $w \in W$, and (ii) that $\mu_i(w)(v) = 0$ for $(w, v) \notin R_i$.

Definition 3.2. An *(enriched) probabilistic Kripke model* is a tuple $\mathbb{M} = \langle \mathbb{F}, V \rangle$, where $\mathbb{F} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2 \rangle$ is an (enriched) probabilistic Kripke frame and $V : \mathit{Prop} \to \wp(W)$ is a valuation.

The probabilistic Kripke models (and frames) defined above are called 'enriched' to distinguish them from the ones used by Fagin and Halpern (1994) and Kooi (2003): our models contain the equivalence relations E_i (whose function will be clarified below), whereas theirs don't. However, the models used in the remainder of this thesis are always the enriched ones defined above; therefore we will henceforth omit the extra qualifier and simply talk about 'probabilistic Kripke models'.

We make some comments on the different components of these models. As usual, R_i is agent i's epistemic accessibility relation: $(w, v) \in R_i$ means that i cannot epistemically distinguish between states w and v. The E_i-relation represents the structure of agent i's experiment: $(w, v) \in E_i$ means that agent i's experiment does not differentiate w and v. Intuitively, we can think of carrying out an experiment as asking a question to nature. This informal analogy carries

over to the formal level: the *experiment relations* E_i play the same role in our framework as the *issue relations* in dynamic epistemic logics of questions (van Benthem and Minica 2009).

The probability mass function $\mu_i(w)$ represents agent i's subjective probabilities (at state w). For example, $\mu_i(w)(v) = a$ means that at state w, agent i assigns subjective probability a to state v being the actual state. The definition of $\mu_i(w)$ is lifted to any set $X \subseteq W$ by putting $\mu_i(w)(X) := \sum_{x \in X} \mu_i(w)(x)$.

We now make some comments on conditions (i) and (ii) of Definition 3.1. Condition (i) is a 'liveness' condition, requiring that the agents do not assign probability 0 to the actual world. At the end of this subsection (after the object language and its semantics have formally been introduced), we will see that this condition corresponds to the principle $p \rightarrow P_i(p) > 0$, i.e. the agents assign non-zero probability to truths. This principle thus requires the agents to be 'prudent': if an agent doesn't know whether p is true, then, to make sure she's complying with this principle, she should assign non-zero probability to both p and $\neg p$. The main reasons for including this condition are, however, of a more technical nature. In the next subsections we will introduce several ways of updating probabilistic Kripke models, all of which say that the agents' probabilities change via Bayesian conditionalization. This requires, however, that $\mu_i(w)(X) > 0$ for several sets $X \subseteq W$. Condition (i) is an easy way to ensure that $\mu_i(w)(X) > 0$ for all the relevant sets X. Finally, note that an assumption similar to condition (i) is also made by Aumann himself.[3]

Condition (ii) says that the agents have to assign probability 0 to all states that they can epistemically distinguish from the actual state (i.e. that they know not to be the actual state). At the end of this subsection, we will see that this condition corresponds to the principle $K_i p \rightarrow P_i(p) = 1$, i.e. the agents assign probability 1 to all the propositions that they know. This seems to be a very reasonable demand for rational agents. Technically speaking, condition (ii) leads to the following easy, but very useful lemma.[4]

Lemma 1. *Let* $\mathbb{M} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2, V \rangle$ *be an arbitrary probabilistic Kripke model and* $w \in W$*. For any set* $X \subseteq W$ *it holds that* $\mu_i(w)(X \cap R_i[w]) = \mu_i(w)(X)$.

We now introduce the (static) language $\mathcal{L}$ by means of the following Backus-

[3] Literally: "$\mathcal{P}_1$ and $\mathcal{P}_2$ [are] partitions of Ω whose join $\mathcal{P}_1 \vee \mathcal{P}_2$ *consists of nonnull events*" (Aumann 1976, p. 1236, my emphasis).

[4] For any binary relation $\mathcal{R} \subseteq W \times W$, we abbreviate $\mathcal{R}[w] := \{v \in W \mid (w, v) \in \mathcal{R}\}$. Furthermore, we will write $\mathcal{R}^*$ for the reflexive transitive closure of $\mathcal{R}$ and $\mathcal{R}^+$ for the transitive closure of $\mathcal{R}$. Finally, note that for reasons of space, most proofs have been omitted from this paper. All details can be found in Demey (2010).

Naur form:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_i\varphi \mid R_i\varphi \mid C^{\varphi}\varphi \mid X^{\varphi}\varphi \mid a_1P_i(\varphi_1) + \cdots + a_nP_i(\varphi_n) \geq k$$

(where $i \in \{1, 2\}$, $1 \leq n < \omega$ and $a_1, \ldots, a_n, k \in \mathbb{Q}$). We only allow rational numbers as values for $a_1, \ldots, a_n, k$ in order to keep the language countable.

As usual, $K_i\varphi$ means that agent i knows that φ. Furthermore, we have the *relativized common knowledge* operator $C^{\varphi}\psi$, which intuitively says that if φ is announced, then it becomes common knowledge (among agents 1 and 2) that ψ was the case before the announcement. The reason for introducing a relativized instead of an ordinary common knowledge operator is well-known: because of its higher expressivity, relativized common knowledge allows for the formulation of a reduction axiom under public announcements (van Benthem et al. 2006).

Knowledge and (relativized) common knowledge have 'post-experimental' counterparts: $R_i\varphi$ and $X^{\varphi}\psi$.[5] First, $R_i\varphi$ says that after carrying out the experiments, agent i will know that φ was the case before the experiments. Second, $X^{\varphi}\psi$ says that after carrying out the experiments, if φ is announced, then it becomes common knowledge (among agents 1 and 2) that ψ was the case before the experiments and the announcement. These operators 'pre-encode' the effects of the experiments in the static language, and will thus enable us to express reduction axioms for the dynamic experimentation operator that will be introduced in the next subsection.[6]

Formulas of the form $a_1P_i(\varphi_1) + \cdots + a_nP_i(\varphi_n) \geq k$ will be called *i-probability formulas*. Note that we do not allow mixed agent indices in such formulas; e.g. $P_1(\varphi) + P_2(\psi) \geq k$ is *not* a well-formed formula. Intuitively, $P_i(\varphi) \geq k$ says that agent i assigns probability at least k to φ. There are two reasons for allowing summation and multiplication by rationals: (i) this extra expressivity is useful when establishing completeness results, and (ii) more importantly, it allows us to express comparative judgments such as 'agent i thinks that φ is at least twice as probable as ψ': $P_i(\varphi) \geq 2P_i(\psi)$.[7]

[5]Hence we have two R_i's: on the one hand, R_i is agent i's epistemic accessibility relation in a probabilistic Kripke model $\mathbb{M}$; on the other hand, R_i is a unary modal operator of the language $\mathcal{L}$. Our main reason for not using another letter for the post-experimental knowledge operator is to ensure uniformity of notation with van Benthem and Minica (2009). We trust that the meaning of R_i will always be clear from the context.

[6]Ordinary (post-experimental) common knowledge can be defined as $C\varphi := C^{\top}\varphi$ and $X\varphi := X^{\top}\varphi$. Furthermore, we define (post-experimental) general knowledge by putting $E\varphi := K_1\varphi \wedge K_2\varphi$ and $F\varphi := R_1\varphi \wedge R_2\varphi$.

[7]This last formula is actually an abbreviation for $P_i(\varphi) - 2P_i(\psi) \geq 0$. One easily sees that the

Consider a probabilistic Kripke model $\mathbb{M} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2, V\rangle$ and a state $w \in W$. Now and in the remainder of this paper, we will often abbreviate $R := R_1 \cup R_2$, $R^e := (R_1 \cap E_1) \cup (R_2 \cap E_2)$, and $[\![\varphi]\!]^{\mathbb{M}} := \{v \in W \mid \mathbb{M}, v \models \varphi\}$. The formal semantics of $\mathcal{L}$ is inductively defined as follows:

$$
\begin{array}{lll}
\mathbb{M}, w \models p & \text{iff} & w \in V(p) \\
\mathbb{M}, w \models \neg\varphi & \text{iff} & \mathbb{M}, w \not\models \varphi \\
\mathbb{M}, w \models \varphi \wedge \psi & \text{iff} & \mathbb{M}, w \models \varphi \text{ and } \mathbb{M}, w \models \psi \\
\mathbb{M}, w \models K_i\varphi & \text{iff} & \forall v \in W : (w, v) \in R_i \Rightarrow \mathbb{M}, v \models \varphi \\
\mathbb{M}, w \models C^{\varphi}\psi & \text{iff} & \forall v \in W : (w, v) \in \Big(R \cap (W \times [\![\varphi]\!]^{\mathbb{M}})\Big)^{+} \Rightarrow \mathbb{M}, v \models \psi \\
\mathbb{M}, w \models R_i\varphi & \text{iff} & \forall v \in W : (w, v) \in R_i \cap E_i \Rightarrow \mathbb{M}, v \models \varphi \\
\mathbb{M}, w \models X^{\varphi}\psi & \text{iff} & \forall v \in W : (w, v) \in \Big(R^e \cap (W \times [\![\varphi]\!]^{\mathbb{M}})\Big)^{+} \Rightarrow \mathbb{M}, v \models \psi \\
\mathbb{M}, w \models \sum_{\ell=1}^{n} a_\ell P_i(\varphi_\ell) \geq k & \text{iff} & \sum_{\ell=1}^{n} a_\ell \mu_i(w)([\![\varphi_\ell]\!]^{\mathbb{M}}) \geq k
\end{array}
$$

Truth and validity at models, frames, and classes of frames are defined as usual. As promised earlier, we finish this subsection with frame correspondence results for conditions (i) and (ii) of Definition 3.1.

Lemma 2. *Let $\mathbb{F} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2\rangle$ be an arbitrary probabilistic Kripke frame. Then we have:*

1. *for all $w \in W : \mu_i(w)(w) > 0$* *iff* $\mathbb{F} \models p \rightarrow P_i(p) > 0$
2. *for all $w, v \in W$: if $(w, v) \notin R_i$ then $\mu_i(w)(v) = 0$* *iff* $\mathbb{F} \models K_i p \rightarrow P_i(p) = 1$

3.2 Dynamics: the experimentation phase

We will now model the first type of dynamics described in Section 2, viz. carrying out the experiments. Syntactically, we add a new dynamic operator [EXP] to the language $\mathcal{L}$, thus obtaining the language $\mathcal{L}([\text{EXP}])$. The [EXP]-operator says that both agents perform their experiments; hence, $[\text{EXP}]\,\varphi$ is to be read as: 'after the agents have performed their experiments, φ holds'. The semantic clause for the [EXP]-operator involves going from the model $\mathbb{M}$ to the updated model $\mathbb{M}^e$, which is defined immediately afterwards.

$$\mathbb{M}, w \models [\text{EXP}]\,\varphi \;\text{ iff }\; \mathbb{M}^e, w \models \varphi$$

format of *i*-probability formulas is sufficiently general to express any 'equation' concerning *i*'s probabilities, cf. Fagin and Halpern (1994).

Definition 3.3. Let $\mathbb{M} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2, V\rangle$ be an arbitrary probabilistic Kripke model. The updated model $\mathbb{M}^e = \langle W^e, R_1^e, R_2^e, E_1^e, E_2^e, \mu_1^e, \mu_2^e, V^e\rangle$ is defined as follows:

- $W^e := W$, $R_i^e := R_i \cap E_i$, and $E_i^e := E_i$
- for all $w \in W^e$, put $\mu_i^e(w)\colon W^e \to [0,1]\colon v \mapsto \mu_i^e(w)(v) := \frac{\mu_i(w)(\{v\}\cap E_i[w])}{\mu_i(w)(E_i[w])}$
- for all $p \in Prop$, put $V^e(p) := V(p)$

Recall that we abbreviated $R^e = (R_1 \cap E_1) \cup (R_2 \cap E_2)$ in the previous section. Applying Definition 3.3, this can now be rewritten as $R^e = R_1^e \cup R_2^e$, which is structurally analogous to our other abbreviation: $R = R_1 \cup R_2$.

We will now justify our definition of the model update operation $\mathbb{M} \mapsto \mathbb{M}^e$ by explaining the intuitions behind it, and by showing that it leads to the right results in a concrete scenario. Carrying out the experiments does not change the set of possible states. Experiment 1 intersects agent 1's accessibility relation R_1 with the experiment relation E_1, and leaves agent 2's accessibility relation unchanged. Symmetric remarks hold for experiment 2.[8] This closely resembles the description by Bonanno and Nehring (1997) of the experiments as imposing a partition on the model.

We now turn to the probabilistic component. The definition of $\mu_i^e(w)$ can be rewritten in terms of conditional probabilities: $\mu_i^e(w)(x) = \mu_i(w)(x \mid E_i[w])$; i.e. agent i conditionalizes on the information that she has gained by performing her experiment. This captures the idea that the agents process new information by means of Bayesian updating.[9]

Example 1. Consider the following scenario. Agent 1 does not know whether p is the case, i.e. she cannot distinguish between p-states and $\neg p$-states. (At the actual state, p is true.) Furthermore, agent 1 has no specific reason to think that one state is more probable than any other; therefore it is reasonable for her to assign equal probabilities to all states. Finally, although agent 1 does not know whether p is the case, she has an experiment that discriminates between p-states and $\neg p$-states, and that thus, when carried out, will allow her to find out whether p is the case. (Agent 2 does not play a role in this scenario.)

[8] We already discussed the analogy between carrying out an experiment and asking a question. Our modeling of the experiments as intersecting R_i with E_i is analogous to the 'resolve' action in the dynamic epistemic logic of questions (cf. van Benthem and Minica (2009, Definition 6)): carrying out an experiment means getting an answer to a question posed to nature.

[9] Note that μ_i^e is well-defined (no 0-divisions): since E is an equivalence relation, it holds that $w \in E_i[w]$, so by condition (i) in Definition 3.1 it follows that $\mu_i(w)(E_i[w]) \geq \mu_i(w)(w) > 0$.

Consider the model $\mathbb{M} := \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2, V\rangle$, with $W := \{w, v\}$, $R_1 := W \times W$, $E_1 = \{(w,w),(v,v)\}$, $\mu_1(w)(w) = \mu_1(w)(v) = \frac{1}{2}$, and $V(p) = \{w\}$ (we do not care about the definitions of $\mu_1(v), R_2, E_2$ and μ_2). It is easy to see that this model is a faithful representation of the above scenario. Consider, for example:

$$\mathbb{M}, w \models \neg K_1 p \wedge \neg K_1 \neg p \wedge P_1(p) = \frac{1}{2} \wedge P_1(\neg p) = \frac{1}{2}$$

Now suppose that the agents carry out their experiments, i.e. consider the updated model $\mathbb{M}^e$. Applying Definition 3.3, it is easy to see that

$$\mathbb{M}, w \models [\text{EXP}]\Big(K_1 p \wedge P_1(p) = 1 \wedge P_1(\neg p) = 0\Big)$$

So after carrying out her experiment, agent 1 has come to know that p is in fact the case. She has also adjusted her probabilities: she now assigns probability 1 to p being true, and probability 0 to p being false. These are the results that we would expect intuitively. Therefore, Definition 3.3 seems to be a natural way of representing the experimentation dynamics: it makes the intuitively right 'predictions' about the agents' knowledge and probabilities.

3.3 Dynamics: the communication phase

We will now model the second type of dynamics described in Section 2, viz. the communication phase. Informally, we treat the communication as a *dialogue about* φ, i.e. a sequence in which the agents each repeatedly communicate the subjective probability they assign to φ (at that point in the dialogue). Single steps in the dialogue are modeled as public announcements.

Public announcements

We first introduce single public announcements. Syntactically, we add a new dynamic operator $[!\cdot]$ to the language $\mathcal{L}([\text{EXP}])$, thus obtaining the language $\mathcal{L}([\text{EXP}], [!\cdot])$. The public announcement operator $[!\varphi]$ says that the formula φ is truthfully and publicly announced to all agents. Hence, $[!\varphi]\psi$ is to be read as: 'after the truthful public announcement of φ, it will be the case that ψ'. The truthfulness of the announcement is captured by means of a precondition in the semantic clause:

$$\mathbb{M}, w \models [!\varphi]\psi \text{ iff } (\text{if } \mathbb{M}, w \models \varphi \text{ then } \mathbb{M}^{\varphi}, w \models \psi)$$

Definition 3.4. Let $\mathbb{M} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2, V\rangle$ be an arbitrary probabilistic Kripke model and $\varphi \in \mathcal{L}([\text{EXP}], [!\cdot])$ an arbitrary formula. The updated model $\mathbb{M}^\varphi = \langle W^\varphi, R_1^\varphi, R_2^\varphi, E_1^\varphi, E_2^\varphi, \mu_1^\varphi, \mu_2^\varphi, V^\varphi\rangle$ is defined as follows:

- $W^\varphi := [\![\varphi]\!]^{\mathbb{M}} = \{w \in W \mid \mathbb{M}, w \models \varphi\}$
- $R_i^\varphi := R_i \cap ([\![\varphi]\!]^{\mathbb{M}} \times [\![\varphi]\!]^{\mathbb{M}})$ and $E_i^\varphi := E_i \cap ([\![\varphi]\!]^{\mathbb{M}} \times [\![\varphi]\!]^{\mathbb{M}})$
- for all $w \in W^\varphi$, put $\mu_i^\varphi(w)\colon W^\varphi \to [0,1]\colon v \mapsto \mu_i^\varphi(w)(v) := \frac{\mu_i(w)(\{v\} \cap [\![\varphi]\!]^{\mathbb{M}})}{\mu_i(w)([\![\varphi]\!]^{\mathbb{M}})}$
- for all $p \in \mathit{Prop}$, put $V^\varphi(p) := V(p) \cap [\![\varphi]\!]^{\mathbb{M}}$

We will now justify our definition of the model update operation $\mathbb{M} \mapsto \mathbb{M}^\varphi$ by showing that it nicely captures the intuitive idea of the public announcement of a formula φ. As usual, the main effect of the public announcement of φ is that all $\neg\varphi$-states get deleted. The other components, R_i, E_i and V, change accordingly. We now turn to the probabilistic component. The definition of $\mu_i^\varphi(w)$ can be rewritten in terms of conditional probabilities: $\mu_i^\varphi(w)(x) = \mu_i(w)(x \mid [\![\varphi]\!]^{\mathbb{M}})$; i.e. agent i conditionalizes on (the information conveyed by) the formula that was publicly announced.[10] This idea can also be expressed in the object language, by means of the following formula (cf. Kooi (2003, p. 394)):[11]

$$\varphi \longrightarrow \Big([!\varphi]P_i(\psi) = k \leftrightarrow P_i([!\varphi]\psi \mid \varphi) = k\Big)$$

It is easy to check that this formula is true on all probabilistic Kripke models. The antecedent mentions the truthfulness precondition of public announcements. The consequent says that public announcement is related to Bayesian conditionalization (modulo dynamic effects): agent i's probability for ψ *after* the public announcement of φ is the same as her probability *before* the announcement for $[!\varphi]\psi$, conditional on φ.

Definition 3.4 fits well with our intuitive idea of what a public announcement of φ is, and how it influences the agents' knowledge and probabilities. One can easily construct scenarios similar to Example 1, in which the 'predictions' about the agents' knowledge and probabilities made by Definition 3.4 match perfectly with our intuitive expectations.

[10] Note that μ_i^φ is well-defined (no 0-divisions): $\mu_i^\varphi(w)$ is only defined for states $w \in W^\varphi = [\![\varphi]\!]^{\mathbb{M}}$, so by condition (i) of Definition 3.1 it follows that $\mu_i(w)([\![\varphi]\!]^{\mathbb{M}}) \geq \mu_i(w)(w) > 0$.

[11] We use $P_i([!\varphi]\psi \mid \varphi) = k$ as an abbreviation for $P_i(\varphi \wedge [!\varphi]\psi) = kP_i(\varphi)$.

Dialogues

We will now move from *single* public announcements to *sequences* of public announcements. We will focus on one particular type of such sequences, which will be called a *dialogue about* φ. In a dialogue about φ, each agent repeatedly announces the probability she assigns to φ (at that step in the dialogue). We will show that such dialogues reach a fixed point after finitely many steps.

Consider a probabilistic Kripke model $\mathbb{M} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2, V\rangle$, a state $w \in W$ and a formula φ. Note that there are unique $a, b \in \mathbb{R}$ such that $\mu_1(w)([\![\varphi]\!]^{\mathbb{M}}) = a$ and $\mu_2(w)([\![\varphi]\!]^{\mathbb{M}}) = b$; i.e. such that $\mathbb{M}, w \models P_1(\varphi) = a \wedge P_2(\varphi) = b$. We now define the sentence $d(\mathbb{M}, w, \varphi)$ as follows:[12]

$$d(\mathbb{M}, w, \varphi) := P_1(\varphi) = a \wedge P_2(\varphi) = b$$

Note that for any model $\mathbb{M}$, state w of $\mathbb{M}$, and formula φ, it holds —by definition of $d(\mathbb{M}, w, \varphi)$— that

$$\mathbb{M}, w \models d(\mathbb{M}, w, \varphi) \tag{1}$$

A single step in the dialogue consists of both agents publicly announcing the probabilities they assign to φ (at that point in the dialogue). In other words, a single step consists of the public announcement of the sentence $P_1(\varphi) = a \wedge P_2(\varphi) = b$, for the unique $a, b \in \mathbb{R}$ that make this sentence true.

For any probabilistic Kripke model $\mathbb{M}$ that contains w, we define $f_{w,\varphi}(\mathbb{M})$ to be the result of publicly announcing the sentence $d(\mathbb{M}, w, \varphi)$ in the model $\mathbb{M}$ (cf. Definition 3.4). Formally: $f_{w,\varphi}(\mathbb{M}) := \mathbb{M}^{d(\mathbb{M},w,\varphi)}$. It is easy to see that it makes sense to reiterate $f_{w,\varphi}$, i.e. that expressions such as $f^n_{w,\varphi}(\mathbb{M})$ make sense for all $n \geq 1$. Consider, for example, a probabilistic Kripke model $\mathbb{M}$ that contains the state w. Unraveling the definitions, we see that

$$f^2_{w,\varphi}(\mathbb{M}) = f_{w,\varphi}(f_{w,\varphi}(\mathbb{M})) = \left(\mathbb{M}^{d(\mathbb{M},w,\varphi)}\right)^{d(\mathbb{M}^{d(\mathbb{M},w,\varphi)},w,\varphi)}$$

We are now ready to model the entire dialogue about φ, as a sequence in which the agents repeatedly announce the probabilities they assign to φ. Consider a probabilistic Kripke model $\mathbb{M}$ that contains the state w. By repeatedly applying $f_{w,\varphi}$ to $\mathbb{M}$ we obtain a sequence which looks as follows:

$$\mathbb{M} \mapsto f_{w,\varphi}(\mathbb{M}) \mapsto f^2_{w,\varphi}(\mathbb{M}) \mapsto f^3_{w,\varphi}(\mathbb{M}) \mapsto f^4_{w,\varphi}(\mathbb{M}) \mapsto \cdots$$

[12] Note that we have tacitly moved outside the official object language here, because we are writing formulas like $P_1(\varphi) = a \wedge P_2(\varphi) = b$, with *real* numbers a, b, whereas the official object language only contains *rational* numbers. Technically speaking, this can be 'repaired' (cf. Demey (2010)), and it does not matter from a modeling perspective, so we will not dwell on it further.

The following lemma says that the models in this sequence do not continue to change ad infinitum, i.e. the dialogue reaches a *fixed point* after finitely many steps.

Lemma 3. *Consider a probabilistic Kripke model* $\mathbb{M}$ *that contains the state* w*. Then there exists an* $n \in \mathbb{N}$ *such that* $f^n_{w,\varphi}(\mathbb{M}) = f^{n+1}_{w,\varphi}(\mathbb{M})$*.*

We are now ready to provide an exact definition of the communication dynamics. Syntactically, we add the [DIAL($\cdot$)]-operator to the language $\mathcal{L}([\text{EXP}], [!\,\cdot])$, thus obtaining the language $\mathcal{L}([\text{EXP}], [!\,\cdot], [\text{DIAL}(\,\cdot\,)])$ (this is the final, and most expressive, language considered in this paper). The [DIAL(φ)]-operator says that both agents carry out a dialogue about φ, i.e. they repeatedly announce the probabilities they assign to φ, until a fixed point is reached (Lemma 3 guarantees that such a fixed point will indeed always be reached after finitely many steps). Hence, [DIAL(φ)]ψ is to be read as: 'after the agents have carried out a dialogue about φ, it will be the case that ψ'.

The semantic clause for [DIAL(φ)] involves going to the fixed point model $\mathbb{M}^{dial_w(\varphi)}$, which is defined immediately afterwards.

$$\mathbb{M}, w \models [\text{DIAL}(\varphi)]\psi \;\text{ iff }\; \mathbb{M}^{dial_w(\varphi)}, w \models \psi$$

Definition 3.5. Let $\mathbb{M} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2, V\rangle$ be an arbitrary probabilistic Kripke model, $w \in W$ an arbitrary state, and φ an arbitrary formula. Then we define $\mathbb{M}^{dial_w(\varphi)} := f^n_{w,\varphi}(\mathbb{M})$ —where n is the least number such that $f^n_{w,\varphi}(\mathbb{M}) = f^{n+1}_{w,\varphi}(\mathbb{M})$ (this number is guaranteed to exist, because of Lemma 3).

Observation 1. Recall that we assume public announcements to be *truthful*. Furthermore, we have modeled a dialogue about φ as a sequence of public announcements. However, the semantics of [DIAL(φ)] does not involve any preconditions. The reason for this is that the formulas being announced throughout the sequence are true *by definition*, cf. (1). Because a dialogue about φ always takes on this form (it will never involve the announcement of other formulas than $d(\mathbb{K}, w, \varphi)$, for probabilistic Kripke models $\mathbb{K}$), the truth precondition can safely be left out.

4 Agreement Theorems in PDEL

Using the semantic setup introduced in the previous section, we will now formulate and prove various dynamic agreement theorems in probabilistic

dynamic epistemic logic. In Section 4.1 we discuss agreement theorems that make the experimentation dynamics explicit, but still leave the communication implicit. In Section 4.2 we build on this and formulate agreement theorems that make both the experimentation *and* the communication dynamics explicit.

4.1 Agreement theorems in PDEL: only experimentation

Before turning to the first agreement theorem in probabilistic dynamic epistemic logic, we formulate two easy auxiliary lemmas.

Lemma 4. *Let* $\mathbb{M} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2, V\rangle$ *be an arbitrary probabilistic Kripke model and* $w \in W$ *a state of* $\mathbb{M}$*. Then for* $i = 1, 2$*, the set* $R^*[w]$ *can be finitely partitioned in cells of the form* $R_i[v_\ell]$*; i.e. it can be expressed as* $R^*[w] = R_i[v_1] \cup \cdots \cup R_i[v_m]$*, with all the* $R_i[v_\ell]$ *pairwise disjoint.*

Lemma 5. *Let* $\mathbb{M} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2, V\rangle$ *be an arbitrary probabilistic Kripke model and* $w \in W$ *a state of* $\mathbb{M}$*. Consider sets* $X, Y \subseteq W$ *and a partition* $\{Y_1, \ldots, Y_m\}$ *of* Y*. Furthermore, assume that for each element* Y_ℓ *of the partition it holds that* $\mu_i(w)(Y_\ell) > 0$ *and that* $\frac{\mu_i(w)(X\cap Y_\ell)}{\mu_i(w)(Y_\ell)} = a$*. Then also* $\mu_i(w)(Y) > 0$ *and* $\frac{\mu_i(w)(X\cap Y)}{\mu_i(w)(Y)} = a$*.*

We are now ready to formulate and prove the first agreement theorem for probabilistic dynamic epistemic logic:

Theorem 1. *Let* $\mathbb{M} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2, V\rangle$ *be an arbitrary probabilistic Kripke model and* $w \in W$ *a state of* $\mathbb{M}$*. Suppose that the following conditions hold:*

(1) $\mu_1(w) = \mu_2(w)$

(2) for all $v \in R^*[w]$: $\mu_i(w) = \mu_i(v)$

Then we have:

$$\mathbb{M}, w \models [\mathrm{EXP}]\, C(P_1(\varphi) = a \wedge P_2(\varphi) = b) \rightarrow a = b$$

Proof. Assume that $\mathbb{M}, w \models [\mathrm{EXP}]\, C(P_1(\varphi) = a \wedge P_2(\varphi) = b)$; we will show that $\mathbb{M}, w \models a = b$, i.e. simply that $a = b$.

Applying Lemma 4 to $\mathbb{M}^e$ (for agent 1), we express $(R^e)^*[w] = R_1^e[v_1] \cup \cdots \cup R_1^e[v_m]$, with all the $R_1^e[v_\ell]$ pairwise disjoint. Now consider any ℓ between 1 and m. Since R_1^e is reflexive, we have $v_\ell \in R_1^e[v_\ell] \subseteq (R^e)^*[w]$. Since $\mathbb{M}, w \models [\mathrm{EXP}]\, C(P_1(\varphi) = a \wedge P_2(\varphi) = b)$, we get $\mathbb{M}^e, w \models C(P_1(\varphi) = a \wedge P_2(\varphi) = b)$, so $v_\ell \in (R^e)^*[w]$ implies that $\mathbb{M}^e, v_\ell \models P_1(\varphi) = a \wedge P_2(\varphi) = b$. Hence $\mu_1^e(v_\ell)([\![\varphi]\!]^{\mathbb{M}^e}) = a$ (†). Note that $R^e = (R_1 \cap E_1) \cup (R_2 \cap E_2) \subseteq R_1 \cup R_2 = R$, and hence $v \in (R^e)^*[w] \subseteq R^*[w]$, so condition 2 of this theorem applies to v_ℓ, i.e. $\mu_1(w) = \mu_1(v_\ell)$ (‡). We now have:

$$\begin{aligned} a &= \mu_1^e(v_\ell)([\![\varphi]\!]^{\mathbb{M}^e}) && (\dagger)\\ &= \frac{\mu_1(v_\ell)([\![\varphi]\!]^{\mathbb{M}^e} \cap E_1[v_\ell])}{\mu_1(v_\ell)(E_1[v_\ell])} && (\text{Def. 3.3})\\ &= \frac{\mu_1(v_\ell)([\![\varphi]\!]^{\mathbb{M}^e} \cap E_1[v_\ell] \cap R_1[v_\ell])}{\mu_1(v_\ell)(E_1[v_\ell] \cap R_1[v_\ell])} && (\text{Lemma 1})\\ &= \frac{\mu_1(w)([\![\varphi]\!]^{\mathbb{M}^e} \cap R_1^e[v_\ell])}{\mu_1(w)(R_1^e[v_\ell])} && (\ddagger) \end{aligned}$$

(Note that $\mu_1(w)(R_1^e[v_\ell]) = \mu_1(v_\ell)(R_1[v_\ell] \cap E_1[v_\ell]) = \mu_1(v_\ell)(E_1[v_\ell]) > 0$.) As ℓ was chosen arbitrarily, this holds for all $1 \leq \ell \leq m$. By Lemma 5 it now follows that $\mu_1(w)\Big((R^e)^*[w]\Big) > 0$ and

$$\frac{\mu_1(w)\Big([\![\varphi]\!]^{\mathbb{M}^e} \cap (R^e)^*[w]\Big)}{\mu_1(w)\Big((R^e)^*[w]\Big)} = a \tag{2}$$

It is easy to see that the entire argument presented above can also be carried out for agent 2. The conclusion of this second, analogous argument will be that

$$\frac{\mu_2(w)\Big([\![\varphi]\!]^{\mathbb{M}^e} \cap (R^e)^*[w]\Big)}{\mu_2(w)\Big((R^e)^*[w]\Big)} = b \tag{3}$$

Now recall condition 1 of this theorem: $\mu_1(w) = \mu_2(w)$. Hence (2) and (3) together imply that $a = b$. □

Observation 2. The reader familiar with Aumann (1976) will probaby have noticed that the proof of our agreement theorem in probabilistic dynamic epistemic logic is a straightforward adaptation of Aumann's own proof for his original agreement theorem (but incorporating already the experimentation dynamics, whereas Aumann's theorem is fully static; cf. Subsection 6.2). This shows that probabilistic Kripke models are a natural setting in which to formalize (dynamic) agreement theorems.

We will now comment on the intuitive interpretation of this theorem and on the two assumptions required to prove it. The theorem is essentially a sentence of the formal language $\mathcal{L}([\text{EXP}])$, and says that if after carrying out the experiments, the agents reach common knowledge about their posteriors for φ, then these posteriors have to be identical. Intuitively, this is very close to Aumann's original agreement theorem, but with the experimentation dynamics explicitly represented in the language. Note, however, that this theorem says what will be the case *if* the agents reach common knowledge for their posterior

about φ, without saying anything about *how* such common knowledge is to be achieved.

The two conditions required to prove the agreement theorem are fairly weak. Condition 1 ($\mu_1(w) = \mu_2(w)$) is an immediate formalization of Aumann's 'common prior' assumption, but localized to the concrete state w. Condition 2 ($\mu_i(w) = \mu_i(v)$ for all $v \in R^*[w]$) is a weakened version of an assumption that is also implicit in Aumann's original setup: Aumann works with structures which have just *one* probability mass function, i.e. he assumes that $\mu_i(x) = \mu_i(y)$ for all states $x, y \in W$. Our theorem shows that this assumption can be weakened: the local version ($\mu_i(x) = \mu_i(w)$ for all $x \in R^*[w]$) suffices. In Subsection 5.2 we will show that common knowledge is not needed to characterize this property: individual knowledge suffices.

It should be noted that Theorem 1 is a *local* theorem (about a particular state w) and a theorem about probabilistic Kripke *models*. However, in the proof we nowhere made any use of the concrete valuation. Furthermore, also the reference to the concrete state w can be eliminated by 'de-localizing' the theorem's two assumptions. In this way, we arrive at the following *global frame version* of the first agreement theorem:

Theorem 2. *Let $\mathbb{F} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2 \rangle$ be an arbitrary probabilistic Kripke frame. Suppose that the following conditions hold:*

(1) $\mu_1 = \mu_2$

(2) for all $w, v \in W$: if $(w, v) \in R^*$ then $\mu_i(w) = \mu_i(v)$

Then we have:

$$\mathbb{F} \models [\mathrm{EXP}]\, C(P_1(\varphi) = a \wedge P_2(\varphi) = b) \rightarrow a = b$$

4.2 Agreement theorems in PDEL: experimentation *and* communication

We now turn to the second agreement theorem in probabilistic dynamic epistemic logic, which also explicitly represents the communication dynamics (in contrast with the first agreement theorem).

First, however, we need to prove one more auxiliary lemma. Intuitively, this lemma says that after a dialogue about φ, the agents' probabilities for φ have become common knowledge.

Lemma 6. *Let* $\mathbb{M} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2, V\rangle$ *be an arbitrary probabilistic Kripke model and assume that* $w \in W$*. Then*

$$\mathbb{M}, w \models [\mathrm{DIAL}(\varphi)]\Big((P_1(\varphi) = a \wedge P_2(\varphi) = b) \rightarrow C(P_1(\varphi) = a \wedge P_2(\varphi) = b)\Big)$$

We are now ready to formulate and prove the second agreement theorem for probabilistic dynamic epistemic logic, which explicitly represents both the experimentation and the communication dynamics:

Theorem 3. *Let* $\mathbb{M} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2, V\rangle$ *be an arbitrary probabilistic Kripke model and* $w \in W$ *a state of* $\mathbb{M}$*. Suppose that the following conditions hold:*

(1) $\mu_1(w) = \mu_2(w)$

(2) for all $v \in R^*[w] : \mu_i(w) = \mu_i(v)$

Then we have:

$$\mathbb{M}, w \models [\mathrm{EXP}]\,[\mathrm{DIAL}(\varphi)](P_1(\varphi) = a \wedge P_2(\varphi) = b) \rightarrow a = b$$

Proof. This proof is structurally analogous to that of Theorem 1, but it makes an essential use of Lemma 6. □

The theorem says that after the agents have carried out the experiments, and then carried out a dialogue about φ, their posteriors for φ have to be identical. Intuitively, this is very close to Aumann's original agreement theorem, except that the experimentation and communication dynamics are now explicitly represented in the language.

Observation 3. In the first agreement theorem, we said that if the agents have common knowledge of their posteriors, then these posteriors have to be identical. However, we said nothing about how this common knowledge is to be achieved, i.e. we did not say anything about the communication. Now, however, we *do* explicitly represent the communication dynamics, and we thus no longer need the common knowledge operator in the formulation of the theorem: the existence of common knowledge can now be *derived* as the result of the communication (cf. Lemma 6).

We again obtain a *global frame version* of the agreement theorem by 'delocalizing' the assumptions:

Theorem 4. *Let* $\mathbb{F} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2\rangle$ *be an arbitrary probabilistic Kripke frame. Suppose that the following conditions hold:*

(1) $\mu_1 = \mu_2$

(2) for all $w, v \in W$: *if* $(w, v) \in R^*$ *then* $\mu_i(w) = \mu_i(v)$

Then we have:

$$\mathbb{F} \models [\text{EXP}]\,[\text{DIAL}(\varphi)](P_1(\varphi) = a \wedge P_2(\varphi) = b) \rightarrow a = b$$

5 Metatheory

We will now develop a sound and complete logic in which the agreement theorem can be derived. Subsection 5.1 discusses a technical difficulty related to the syntactic perspective on probabilistic epistemic logic in general, and proposes a solution to it. Subsection 5.2 provides characterization results for the conditions of the agreement theorems proved in Section 4. These characterization results are then used in Subsection 5.3 to obtain various axiomatizations.

5.1 A difficulty about expressivity

Our modeling of the experiments has so far been very general: agent i's experiment corresponds to any equivalence relation E_i (or, equivalently, to any partition of the model) whatsoever. From the syntactic perspective, however, this full generality is difficult to maintain, because it exceeds the expressive powers of the formal language $\mathcal{L}([\text{EXP}])$. We will first give a concrete illustration of this problem and then propose a solution to it.

Recall the semantics for i-probability formulas such as $P_i(\varphi) \geq k$:

$$\mathbb{M}, w \models P_i(\varphi) \geq k \quad \text{iff} \quad \mu_i(w)([\![\varphi]\!]^{\mathbb{M}}) \geq k$$

There is a clear asymmetry in expressivity between both sides of this definition. On the left hand side, there is a formula of the formal language $\mathcal{L}([\text{EXP}])$. The Backus-Naur form of this language guarantees that $P_i(\cdot)$ will always receive a formula as its argument. On the right hand side, however, we have the function $\mu_i(w)(\cdot)$, which can receive any set $X \subseteq W$ whatsoever as its argument, even *undefinable* sets (i.e. sets X such that $X = [\![\varphi]\!]^{\mathbb{M}}$ for no $\mathcal{L}([\text{EXP}])$-formula φ). It may well be the case that $E_i[w]$ is an undefinable set. In that case, several problems of expressivity will arise; for example, the [EXP]-reduction axiom will in general not be expressible in $\mathcal{L}([\text{EXP}])$.

To solve the problem we need to make sure that $E_i[w]$ is always definable by means of some formula. One way to ensure this is by restricting to *binary*

experiments.[13] The first, *syntactic* step of this strategy is to introduce two new, 'primitive' formulas α_1, α_2 into the language. The second, *semantic* step involves assuming that for any probabilistic Kripke frame $\mathbb{F} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2 \rangle$ there exist sets $\mathcal{E}_i^{\mathbb{F}} \subseteq W$ such that $E_i = \Big(\mathcal{E}_i^{\mathbb{F}} \times \mathcal{E}_i^{\mathbb{F}}\Big) \cup \Big((W - \mathcal{E}_i^{\mathbb{F}}) \times (W - \mathcal{E}_i^{\mathbb{F}})\Big)$. In the third and final step, we link syntax and semantics by extending the valuations to the newly introduced α_i: for any valuation V on $\mathbb{F}$, we require that $V(\alpha_i) := \mathcal{E}_i^{\mathbb{F}}$, and thus obtain:

$$E_i = \Big(V(\alpha_i) \times V(\alpha_i)\Big) \cup \Big((W - V(\alpha_i)) \times (W - V(\alpha_i))\Big) \tag{4}$$

It is easy to check that E_i, thus defined, is still an equivalence relation, and furthermore, that this new definition is 'compatible' with the main types of dynamics discussed in this paper, in the sense that if a probabilistic Kripke model $\mathbb{M}$ satisfies condition (4), then the updated models $\mathbb{M}^e$ and $\mathbb{M}^\varphi$ will satisfy it as well.

Informally, (4) says that agent i's experiment only differentiates between α_i-states and $\neg\alpha_i$-states; in other words, it is a 'binary experiment'. Continuing the analogy between experiments and questions, carrying out a binary experiment corresponds to asking a yes-no question: 'is α_i the case or not?'.

In this more restricted setup, it follows easily from condition (4) that $E_i[w] = [\![\alpha_i]\!]^{\mathbb{M}}$ if $\mathbb{M}, w \models \alpha_i$, and $E_i[w] = [\![\neg\alpha_i]\!]^{\mathbb{M}}$ otherwise. Hence $E_i[w]$ is now always definable: either by α_i or by $\neg\alpha_i$ (depending on whether $\mathbb{M}, w \models \alpha_i$). This definability result will be used extensively in Subsection 5.3 (in the [EXP]-reduction axiom for i-probability formulas, but also in other axioms).

5.2 Characterization results

In Section 4 we established various dynamic agreement theorems. These theorems required imposing two conditions on probabilistic Kripke models/frames. We will now establish characterization results for (the global frame versions of) these conditions.

We first characterize the common prior assumption, i.e. condition 1 of Theorems 2 and 4. Note that if φ is a 1-probability formula, then we will use $\varphi[P_2/P_1]$ to denote the formula that is obtained by uniformly substituting P_2 for P_1 in φ. It is clear that if φ is a 1-probability formula, then $\varphi[P_2/P_1]$ is a 2-probability

[13] Another solution, based on hybrid logic, is explored in detail in Demey (2010). There it is also argued that the 'binary experiments'-solution is preferable on technical as well as methodological grounds.

formula. Finally, an i-probability formula φ is said to be *atomic* iff it is of the form $\sum_{\ell=1}^{n} a_\ell P_i(p_\ell) \geq k$.

Lemma 7. *Let $\mathbb{F} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2 \rangle$ be an arbitrary probabilistic Kripke frame. Then $\mu_1 = \mu_2$ iff for all atomic 1-probability formulas φ: $\mathbb{F} \models \varphi \leftrightarrow \varphi[P_2/P_1]$.*

We now characterize condition 2 of Theorems 2 and 4:

Lemma 8. *Let $\mathbb{F} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2 \rangle$ be an arbitrary probabilistic Kripke frame. Then we have:*

$$\begin{array}{c} \text{for all } w, v \in W : \text{if } (w,v) \in R^* \text{ then } \mu_i(w) = \mu_i(v) \quad \text{iff} \\ \text{for all atomic } i\text{-probability formulas } \varphi : \mathbb{F} \models \Big(\varphi \rightarrow C\varphi\Big) \wedge \Big(\neg\varphi \rightarrow C\neg\varphi\Big) \end{array}$$

Observation 4. The condition that $\mu_i(w) = \mu_i(v)$ whenever $(w,v) \in R^*$ is a very heavy constraint to impose on probabilistic Kripke frames: it involves the reflexive transitive closure of R, and might therefore be called 'semi-global'. This aspect is reflected in the above characterization result, which makes use of the common knowledge operator C. However, because frame validity is itself a global notion, it is possible to capture the semi-global frame property involving R^* by means of the more modest general knowledge operator E. This result is still not fully satisfactory, however: the principles that $\varphi \rightarrow E\varphi$ and $\neg\varphi \rightarrow E\neg\varphi$ (for atomic i-probability formulas φ) still require the 'public availability' of agent i's subjective probabilistic setup. However, in frames satisfying the common prior property ($\mu_1 = \mu_2$)—and all the frames used to prove the agreement results indeed satisfy this propery—more plausible 'individual' introspection principles suffice: $\varphi \rightarrow K_i\varphi$ and $\neg\varphi \rightarrow K_i\neg\varphi$ (for atomic i-probability formulas φ). Hence, no notion of social (common/general) knowledge is required to characterize the second assumption of the agreement theorems.

Lemma 9. *Let $\mathbb{F} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2 \rangle$ be an arbitrary probabilistic Kripke frame and suppose that $\mu_1 = \mu_2$. Then we have:*

$$\begin{array}{c} \text{for } i = 1,2 \text{ and for all } w, v \in W : \text{if } (w,v) \in R^* \text{ then } \mu_i(w) = \mu_i(v) \quad \text{iff} \\ \text{for } i = 1,2 \text{ and for all atomic } i\text{-probability formulas } \varphi : \\ \mathbb{F} \models \Big(\varphi \rightarrow K_i\varphi\Big) \wedge \Big(\neg\varphi \rightarrow K_i\neg\varphi\Big) \end{array}$$

5.3 The logics

We will now define three logics of increasing strength, and prove them to be sound and complete with respect to natural classes of Kripke frames. The sec-

ond and, especially, the third logic capture the reasoning behind the agreement theorem. For the sake of clarity, these logics are presented in a modular fashion.

The first logic is the basic *probabilistic epistemic logic* PEL, which captures the behavior of the epistemic and probabilistic operators. It does not say anything about agreement theorems. We first give a schematic overview of the logic, and then discuss each of its components separately.

Componentwise axiomatization of PEL
1. the propositional component
2. the individual knowledge component
3. the common knowledge component
4. the linear inequalities component
5. the probabilistic component
6. the pre-/post-experimental interaction component
7. the α_i-component

The propositional component is fully standard, and needs no further comments. The individual knowledge component is also standard, and says that the individual (pre- and post-experimental) knowledge operators K_i and R_i are S5-modal operators. (Note that also Aumann's original result involved S5-type knowledge.) The axioms for relativized common knowledge ($C^{\varphi}\psi$) are immediately adapted from van Benthem et al. (2006), where this notion was introduced. The post-experimental version of relativized common knowledge ($X^{\varphi}\psi$) is governed by the immediate analogues of these axioms. The linear inequalities component component axiomatizes (operations on) linear inequalities of probabilities. This component is mainly of technical use (proving completeness), and is adapted from Fagin and Halpern (1994).

The probabilistic component consists of two parts. The first part is a straightforward formalization of the well-known Kolmogorov axioms of probability; this is also adapted from Fagin and Halpern (1994). The second part consists of the two formulas that characterize properties (i) and (ii) of probabilistic Kripke frames (cf. Definition 3.1 and Lemma 2):

$$\varphi \to P_i(\varphi) > 0 \qquad K_i\varphi \to P_i(\varphi) = 1$$

The intuitive motivation of these principles was already discussed in Section 3. Next, the pre-/post-experimental interaction component describes the influence of the experiments on the agents' (common) knowledge: it says that carrying out the experiments does not make the agents forget anything that they already (commonly) knew before the experiments. Formally:

$$K_i\varphi \to R_i\varphi \qquad C^{\varphi}\psi \to X^{\varphi}\psi$$

The final component of PEL involves the special proposition letters α_i. First of all, there is an axiom which says that the post-experimental knowledge operator R_i can be defined in terms of the usual knowledge operator K_i and these special proposition letters:[14]

$$R_i\varphi \leftrightarrow \Big(\big(\alpha_i \to K_i(\alpha_i \to \varphi)\big) \wedge \big(\neg\alpha_i \to K_i(\neg\alpha_i \to \varphi)\big)\Big)$$

Finally, this component also contains axioms which say that the agents' experiments are *successful*: if α_i is the case, then after carrying out her experiment, agent i will know this, and likewise if α_i is not the case. Formally:

$$\alpha_i \to R_i\alpha_i \qquad \neg\alpha_i \to R_i\neg\alpha_i$$

This concludes the presentation of the basic probabilistic epistemic logic PEL. We now introduce the second logic, viz. *probabilistic epistemic agreement logic* or PEAL. This logic is a simple extension of PEL: one adds an 'agreement component', which consists of the formulas that characterize the two frame properties needed in the agreement theorems (cf. Lemmas 7-9).

Componentwise axiomatization of PEAL		
1–7.	the seven components of PEL	
8.	the agreement component:	
	$\varphi \leftrightarrow \varphi[P_2/P_1]$	(for 1-probability formulas φ)
	$\varphi \to K_i\varphi$ and $\neg\varphi \to K_i\neg\varphi$	(for i-probability formulas φ)

We now introduce the third and final logic, viz. *dynamic probabilistic epistemic agreement logic with explicit experimentation* or DPEALe. This logic is obtained by simply adding the [EXP]-reduction axioms to PEAL.

Componentwise axiomatization of DPEALe	
1–8.	the eight components of PEAL
9.	the reduction axioms for [EXP]

[14] Given this definability result, it might be asked why R_i is still introduced as a *primitive* operator. The reason for doing this is that this operator is only definable *if* we make use of the special proposition letters α_i; we remind the reader that these were only introduced at the beginning of this section, when we shifted from a semantic to a syntactic pespective.

The reduction axioms for [EXP] are displayed below. Most of them are straightforward; we only emphasize the use of R_i to pre-encode the effects of the experimentation dynamics on K_i (similar remarks apply to common knowledge), and the use of α_i in the reduction axiom for i-probability formulas (avoiding non-expressibility, cf. Subsection 5.1).

1. $[\text{EXP}]\, p \quad \leftrightarrow \quad p \qquad (\text{for } p \in \mathit{Prop} \cup \{\alpha_1, \alpha_2\})$
2. $[\text{EXP}]\, \neg\varphi \quad \leftrightarrow \quad \neg\, [\text{EXP}]\, \varphi$
3. $[\text{EXP}](\varphi \wedge \psi) \quad \leftrightarrow \quad [\text{EXP}]\, \varphi \wedge [\text{EXP}]\, \psi$
4. $[\text{EXP}]\, K_i\varphi \quad \leftrightarrow \quad R_i\, [\text{EXP}]\, \varphi$
5. $[\text{EXP}]\, R_i\varphi \quad \leftrightarrow \quad R_i\, [\text{EXP}]\, \varphi$
6. $[\text{EXP}]\, C^{\varphi}\psi \quad \leftrightarrow \quad X^{[\text{EXP}]\varphi}\, [\text{EXP}]\, \psi$
7. $[\text{EXP}]\, X^{\varphi}\psi \quad \leftrightarrow \quad X^{[\text{EXP}]\varphi}\, [\text{EXP}]\, \psi$
8. $[\text{EXP}]\, [\text{EXP}]\, \varphi \quad \leftrightarrow \quad [\text{EXP}]\, \varphi$
9. $[\text{EXP}] \sum_\ell a_\ell P_i(\varphi_\ell) \geq k \quad \leftrightarrow$

$$\begin{cases} \alpha_i \rightarrow \sum_\ell a_\ell P_i([\text{EXP}]\, \varphi_\ell \wedge \alpha_i) \geq k P_i(\alpha_i) \\ \wedge \quad \neg\alpha_i \rightarrow \sum_\ell a_\ell P_i([\text{EXP}]\, \varphi_\ell \wedge \neg\alpha_i) \geq k P_i(\neg\alpha_i) \end{cases}$$

With the logics in place, we now turn towards their soundness and completeness. First we formally introduce the classes of frames with respect to which soundness and completeness results will be proved:

Definition 5.1. We write $\mathcal{PKB}$ for the class of all enriched probabilistic Kripke frames with binary experiments (i.e. satisfying condition (4)).

Definition 5.2. Consider an arbitrary frame $\mathbb{F} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2 \rangle \in \mathcal{PKB}$. Then $\mathbb{F}$ is said to be an *agreement frame* iff it satisfies conditions 1 and 2 from Theorems 2 and 4. We write $\mathcal{AGR}$ for the class of all agreement frames.

Observation 5. We immediately obtain:

(1) $\mathcal{AGR} \models [\text{EXP}]\, C(P_1(\varphi) = a \wedge P_2(\varphi) = b) \rightarrow a = b$

(2) $\mathcal{AGR} \models [\text{EXP}]\, [\text{DIAL}(\varphi)](P_1(\varphi) = a \wedge P_2(\varphi) = b) \rightarrow a = b$

Theorem 5. *We have the following soundness and completeness results:*

(1) The logic PEL *is sound and complete with respect to* $\mathcal{PKB}$.

(2) The logic PEAL *is sound and complete with respect to* $\mathcal{AGR}$.

(3) The logic DPEALe *is sound and complete with respect to* $\mathcal{AGR}$.

Observation 6. Combining this theorem with Observation 5, we immediately get that $\mathsf{DPEALe} \vdash [\mathrm{EXP}]\, C(P_1(\varphi) = a \wedge P_2(\varphi) = b) \rightarrow a = b$. The system DPEALe is thus strong enough to derive a dynamic agreement theorem which explicitly represents the experimentation dynamics.

6 Methodological Comments

In this section we examine some of the methodological and philosophical implications of the technical results established earlier. Subsection 6.1 provides some comments on the role and importance of the notion of common knowledge in agreement results. Subsection 6.2 discusses the issue of static versus dynamic agreement theorems, and argues that the most natural agreement theorems are all dynamic in nature.

6.1 The role of common knowledge

In order to formulate and prove his agreement theorem, Aumann used the notion of *common knowledge*—thus being the first author to introduce this notion in the game-theoretical literature. Therefore, it is widely assumed that common knowledge plays a central role in agreeing to disagree results. Several results established throughout this paper, however, seem to suggest that the importance of common knowledge is not so central as is often thought.

In Aumann's original setup, the (common) prior probability distribution itself is assumed to be common knowledge among the agents. This is reflected in our framework by the characterization results involving $\varphi \rightarrow C\varphi$ (and $\neg\varphi \rightarrow C\neg\varphi$) for i-probability formulas φ. However, we showed that this can be replaced with the much weaker individual probabilistic-epistemic introspection principle $\varphi \rightarrow K_i\varphi$ (for i-probability formulas φ) (cf. Observation 4). In other words, the assumption that the agents' prior probability distributions are common knowledge can be formally captured without making use of the common knowledge operator.

A second observation concerns the role of common knowledge in obtaining consensus (i.e. identical posterior probabilities). Aumann's original theorem says that if after carrying out the experiments, the agents have common knowledge of their posteriors, then these posteriors have to be identical. However, this theorem does not say *how* the agents are to obtain this common knowledge (it just assumes that they have been able to obtain it one way or another). The way to obtain common knowledge is via communication. Once we decide to

make this communication dynamics an explicit part of the story (and thus, to explicitly represent it in the formal language), the notion of common knowledge disappears (cf. Observation 3). Hence, once we decide to represent both the experimentation and the communication dynamics, the agreement theorem can be formulated without making use of the common knowledge operator.

Finally, we remark that our comments on the relative unimportance of common knowledge for agreeing to disagree results are in line with the results by Parikh and Krasucki (1990). They consider groups of more than two agents, in which communication does not occur publicly, but in pairs. They show that, given certain conditions on the communication protocol, the agents will reach *consensus* (identical posteriors), but not common knowledge.

6.2 Static versus dynamic agreement theorems

Aumann's original agreeing to disagree theorem was a static result (cf. Section 2). In this paper we have proved basically two agreement theorems: one in which the experiments are explicitly represented, and one in which both the experiments and the communication are explicitly represented. Hence, all of our agreement theorems are dynamic; we do not have any static agreement theorem at all.

However, the absence of a static (and thus 'classical') agreement theorem is not a disadvantage of our framework. Once one has taken the dynamic turn, it even seems that the only static agreement theorems are rather convoluted. The models that they talk about are chimæras: one such model seems to be composed of 'pieces' taken from many different 'normal' models.

To illustrate this, we focus on the experimentation dynamics. In our approach, we have two clean, 'temporally uniform' models. The model $\mathbb{M} = \langle W, R_1, R_2, E_1, E_2, \mu_1, \mu_2, V\rangle$ represents the agents' knowledge and probabilities *before* the experiments; the model $\mathbb{M}^e = \langle W^e, R_1^e, R_2^e, E_1^e, E_2^e, \mu_1^e, \mu_2^e, V^e\rangle$ represents the agents' knowledge and probabilities *after* the experiments. Now contrast this with Aumann's original agreement theorem. This talks about 'temporally incoherent' models, which represent the agents' knowledge *after* the experiments, but their probability distributions *before* the experiments. Formally, such a chimæric model would be of the form $\langle W, R_1^e, R_2^e, E_1, E_2, \mu_1, \mu_2, V\rangle$: it is obtained by cutting the (temporally uniform) models $\mathbb{M}$ and $\mathbb{M}^e$ into pieces and then pasting these pieces back together in a temporally incoherent way.

This example can be analyzed as follows. The intuitive agreeing to disagree scenario is *intrinsically dynamic* (cf. Section 2). If one wants to prove a static agreement result (like Aumann), then one will need to 'smuggle' this dynamics

in somehow. In our approach, however, all of the dynamics is represented explicitly in the theorems. We have an epistemic plausibility model $\mathbb{M}$ which corresponds to the initial stage (before the experiments), a model $\mathbb{M}^e$ which corresponds to the time immediatey after the experiments, and finally, a model $(\mathbb{M}^e)^{dial_w(\varphi)}$ which corresponds to the final stage after the communication, at which the agents have reached common knowledge of their posteriors. Hence, there exists a complete structural analogy between the intuitive scenario on the one hand and the formal theorem on the other.

7 Conclusion

In this paper we have established various agreement theorems in probabilistic dynamic epistemic logic. In particular, we established model- and frame-based versions of an agreement theorem with experimentation (Theorems 1 and 2), and of an agreement theorem with experimentation and communication (Theorems 3 and 4). We developed a sound and complete logical system within which the first agreement result is derivable syntactically (Theorem 5 and Observation 6). Throughout the paper, we have emphasized our attempts to keep the models and the logics intuitively plausible, and directly connected with Aumann's original agreement result.

We also discussed two methodological implications of these technical results. In the first place, we argued that the role of common knowledge in the agreement theorem is not so important as is often thought. In the second place, we noted that although Aumann's original theorem is a static result, the intuitive motivation behind it contains a lot of dynamics; we then argued that representing this dynamics is essential to obtain a natural agreement result, and that static agreement theorems (including Aumann's original result) are only possible at the expense of a convoluted notion of model. These two considerations are also related with each other. Common knowledge and communication seem to be two sides of the same coin: common knowledge is the result of communication, so if the communication is explicitly represented in the agreement theorem, there is no need anymore to *assume* common knowledge (as this will now *follow* from the communication).

Acknowledgements I presented earlier versions of this paper at various seminars in Amsterdam, Leuven and Tilburg. Thanks to the organizers and the audiences of these talks for their critical questions and comments. In particular, I would like to thank Johan van Benthem, Eric Pacuit, Dick de Jongh, Cédric

Dégremont and Margaux Smets for their detailed feedback. This research was, at various stages, supported by Igor Douven's Formal Epistemology Project, the Huygens Scholarship Programme, and a PhD fellowship of the Fund for Scientific Research – Flanders (FWO).

References

R. Aumann. Agreeing to disagree. *Annals of Statistics*, 4:1236 – 1239, 1976.

G. Bonanno and K. Nehring. Agreeing to disagree: a survey. 1997. Ms.

C. Dégremont and O. Roy. Agreement theorems in dynamic-epistemic logic. In X. He, J. Horty, and E. Pacuit, editors, *Logic, Rationality, and Interaction. LORI 2009 Proceedings*, LNAI 5834, pages 105 – 118. Springer, 2009.

L. Demey. Agreeing to disagree in probabilistic dynamic epistemic logic. Master's thesis, ILLC, Universiteit van Amsterdam, 2010.

R. Fagin and J. Halpern. Reasoning about knowledge and probability. *Journal of the ACM*, 41:340 – 367, 1994.

J. Halpern and Y. Moses. Knowledge and common knowledge in a distributed environment. *Journal of the ACM*, 37:549 – 587, 1990.

B. Kooi. Probabilistic dynamic epistemic logic. *Journal of Logic, Language and Information*, 12:381 – 408, 2003.

D. Lewis. *Convention*. Harvard University Press, 1969.

P. Milgrom and N. Stokey. Information, trade and common knowledge. *Journal of Economic Theory*, 26:1327 – 1347, 1982.

R. Parikh and P. Krasucki. Communication, consensus and knowledge. *Journal of Economic Theory*, 52:178 – 189, 1990.

J. van Benthem and S. Minica. Toward a dynamic logic of questions. In X. He, J. Horty, and E. Pacuit, editors, *Logic, Rationality, and Interaction. LORI 2009 Proceedings*, LNAI 5834, pages 27 – 41. Springer, 2009.

J. van Benthem, J. van Eijck, and B. Kooi. Logics of communication and change. *Information and Computation*, 204:1620 – 1662, 2006.

Dynamified Hybrid Counterfactual Logic

Katsuhiko Sano

JSPS Research Fellow, Department of Humanistic Informatics, Graduate School of Letters, Kyoto University
katsuhiko.sano@gmail.com

Abstract

This paper proposes a dynamification of Sano's (2009) logic, which hybridized David Lewis' counterfactual logic. Our dynamification involves two dynamic modalities, which amount to the public announcement update using link-cutting (Yamada 2006, van Benthem and Liu 2007) and the radical upgrade (van Benthem 2007). Introducing these modalities to Lewis' sphere semantics has two merits. First, if we interpret the underlying domains of sphere models as a set of individuals and things, dynamic modalities enable us to show how we can change the plausibility relation which is relativized to the speaker. Second, our update operation using link-cutting can capture the result of the change of the speaker's attention when producing a sequence of conditionals (Lewis 1973). Moreover, we give reduction axioms to both dynamic modalities.

1 Introduction

Modal logic has broad applicability because we can interpret the domain of Kripke semantics as possible worlds, moments, spatial coordinates, individuals, etc (cf. Blackburn and van Benthem (2007)). Moreover, modal logic handles local propositions well, whose truths depend on the given states. Hybrid logic (cf. Areces and ten Cate (2007)) adds a more expressive power to modal logic, i.e., it allows us to talk about elements of the domain directly in syntax. Since

hybrid syntax (nominals and satisfaction operators) provides a way of handling global propositions whose truths are independent of the given states, we can handle both local and global propositions in hybrid logic.

Sano (2009) proposed a way to combine Lewis' counterfactual logic (1973) and hybrid logic. We interpreted the semantic domain as the set of individuals or things. Let us call this interpretation *description-logic reading* or *egocentric reading*. Sano's combination enables us to regard the following inference as a process of revising a local proposition by using global proposition: 'The pig is Mary. Mary is pregnant. Therefore, the pig is pregnant'. In this example, we regard the sentences containing 'the' as representing a local proposition, since the truth of these sentences depends on who utters them. On the other hand, we regard 'Mary is pregnant' as a global proposition, because its truth is independent of the speaker. To deal with the sentences containing 'the', we make use of Lewis' egocentric reading of these sentences mentioned above (here we need to use the counterfactual connective in terms of description-logic-reading). In order to deal with sentences like the second premise, we use the syntax of hybrid logic.

Lewis (1973) defines the truth condition of counterfactuals based on the semantic structure representing relative closeness, called *comparative similarity*, between possible worlds. In order to deal with the sentences containing 'the', Lewis proposes that his semantic structure represents relative familiarity, called *comparative salience*, between things. Unlike comparative similarity, Lewis claims that comparative salience is easy to change (Lewis 1973, pp.113-7). According to Lewis, one merit of his formalization is that we can capture the following sequence of sentences containing 'the': 'the pig is grunting, the pig with floppy ears is not grunting, and the spotted pig with floppy ears is grunting' (Lewis 1973, p.114). It seems natural to regard that, if we move from the first sentence to the second one, there is a change in the speaker's attention.

However, Lewis does not explicate how the comparative salience changes in formal setting. Modern developments of dynamic epistemic logic enable us to explicate this point. Moreover, dynamic epistemic logic allows us to explain how the speaker's attention changes while producing a sequence of sentences containing 'the'. More precisely, we add to hybrid counterfactual logic two dynamic modalities, which amounts to the public announcement update using link-cutting (Yamada 2006, van Benthem and Liu 2007) and the radical upgrade (van Benthem 2007). Since we concentrate on description-logic reading in our dynamification, we can also regards our study as a dynamification of description logic (Baader and Lutz 2007).

A combination of dynamic epistemic logic and hybrid logic has been pro-

posed by Roy (2009), van Benthem and Minicã (2009), Ulrik Hansen (2011), etc, mainly because this combination gives us the complete axiomatization of the logic of distributed knowledge. All of these studies consider a dynamification in terms of possible-world reading. In this sense, this paper is different from these studies. As mentioned in (van Benthem 2010, Section 13.4), doxastic logic (especially, a logic of conditional belief) and conditional logic have exactly the same semantics. Similarly, our technical study contributes to the combination of dynamic doxastic logic (or dynamic epistemic logic in the broader sense) and hybrid logic.

We proceed as follows. Section 2 first reviews Lewis' semantics for counterfactuals based on sphere models and a motivation for hybridizing Lewis' counterfactual logic. Our motivation is concerned with egocentric or individual interpretation of counterfactual conditionals, that is, we regard the domain of sphere models as a set of individual or things. Moreover, we mention the following logical notions and results from (Sano 2009): a Hilbert-style axiomatization of hybrid counterfactual logic, its completeness and decidability (see Theorem 1), and the notion of bisimulation between sphere models. Section 3 explains our two motivations for a dynamification of hybrid counterfactual logic and then introduce two dynamic modalities, salience update and salience upgrade, which can be regarded as the public announcement update using link-cutting (Yamada 2006, van Benthem and Liu 2007) and the radical upgrade (van Benthem 2007), respectively. We demonstrate that these modalities explain how we can change a given plausibility relation between things or individuals (see Example 2 and Example 4). Our update operation using link-cutting can capture the result of the speaker's attention change when producing a sequence of conditionals (Example 3). Moreover, we show that this update operation respects sphere-bisimulations (see Proposition 4). Since we can give the valid reduction axioms to both dynamic modalities (Propositions 3 and 6), we finally establish that our dynamification of hybrid counterfactual logic will enjoy completeness and decidability (Theorem 2).

2 Static Hybrid Counterfactual Logic

In this section, we review a motivation for Sano's (2009) hybridization of David Lewis' (1973) counterfactual logic in terms of egocentric interpretation of Lewis' sphere models. Then, we also explain some logical notions and results mostly from (Sano 2009) that we will need in the following sections.

2.1 David Lewis' System of Spheres for Counterfactual Logic

Lewis (1973) proposes that the counterfactual conditional $\varphi \boxrightarrow \varphi$ (read 'If it were the case that φ, then it would be the case that ψ') is true at a world w iff $(\varphi \wedge \psi)$-worlds are *relatively closer* to w than $(\varphi \wedge \neg\psi)$-worlds. Lewis define the notion of 'relative closeness' in terms of the following mathematical structure [1].

Definition 2.1. A pair $(W, \$)$ is *a system of spheres* if $W \neq \emptyset$ and $\$: W \to \mathcal{PP}(W)$ satisfies the following (we write '$\$_w$' instead of '$\$(w)$'):

(*S***1)** $\$_w$ *is nested*: If $S, T \in \$_w$, then $S \subseteq T$ or $T \subseteq S$;

(*S***2)** $\$_w$ *is closed under unions*: If $(S_\lambda)_{\lambda\in\Lambda} \subseteq \$_w$, then $\bigcup_{\lambda\in\Lambda} S_\lambda \in \$_w$;

(*S***3)** $\$_w$ *is closed under* (*nonempty*) *intersections*: If $(S_\lambda)_{\lambda\in\Lambda} \subseteq \$_w$ and $\Lambda \neq \emptyset$, then $\bigcap_{\lambda\in\Lambda} S_\lambda \in \$_w$.

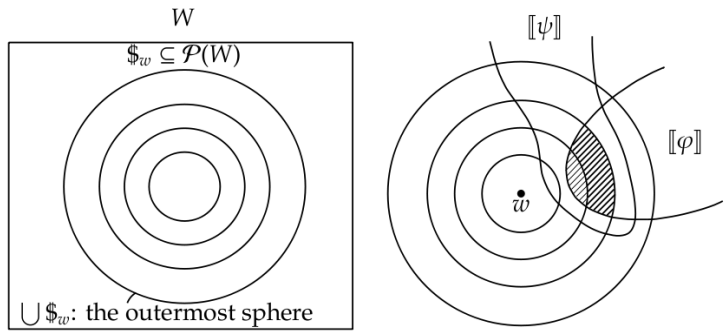


Figure 1: A System of Spheres and the Truth Condition (B) of Counterfactuals

$\bigcup \$_w$ represents the set of all worlds accessible from w. If we identify $\bigcup \$_w$ with $R(w) := \{ w' \in W \mid wRw' \}$ in Kripke semantics for modal logic (where R is a binary relation on W), $(W, \$)$ adds an additional internal structure to $R(w)$.

[1] In order to deal with the *counterfactual conditionals*, Lewis requires the further condition called *Centering*: $\{w\} \in \$_w$ for any $w \in W$. Remark that we assume (Centering) when we deal with counterfactual conditionals, though we will not assume it in general from the next subsection.

This additional structure captures the notion of relative closeness or the notion of *comparative similarity* in Lewis' term.

Remark 2.1. We can define the comparative similarity *relation* $\leq_w$ on $R(w)$ as follows: $u \leq_w v$ iff $v \in S$ implies $u \in S$ for every $S \in \$_w$ (Lewis 1973, p.49). So, we can regard a system $(W, \$)$ of spheres as an (single-agent) *epistemic plausibility model* $(W, R, (\leq_w)_{w \in W})$ (van Benthem 2010, Section 13.3) satisfying the connectedness of $\leq_w$, i.e., $u \leq_w v$ or $v \leq_w u$ for any $u, v \in W$.

Given a valuation V on a system of sphere $(W, \$)$, we can give the truth condition to the counterfactual conditional as follows:

$$(W, \$, V), w \Vdash \varphi \boxdot\!\!\rightarrow \psi \quad \text{iff} \quad \begin{cases} \text{(A) } \bigcup \$_w \cap [\![\varphi]\!] = \emptyset \text{ or} \\ \text{(B) } (\exists S \in \$_w)\,[S \cap [\![\varphi]\!] \neq \emptyset \text{ and } S \cap [\![\varphi]\!] \subseteq [\![\psi]\!]], \end{cases}$$

where $[\![\varphi]\!] := \{ u \in W \mid (W, \$, V), u \Vdash \varphi \}$. It is easy to see that:

$$(W, \$, V), w \Vdash \neg(\varphi \boxdot\!\!\rightarrow \bot) \quad \text{iff} \quad \bigcup \$_w \cap [\![\varphi]\!] \neq \emptyset.$$

By our identification of $\bigcup \$_w$ with '$R(w)$' mentioned above, it is natural to regard $\neg(\varphi \boxdot\!\!\rightarrow \bot)$ as an abbreviation of $\Diamond\varphi$.

The case (B) of the truth condition of $\varphi \boxdot\!\!\rightarrow \psi$ corresponds to our intuitive reading: $(\varphi \wedge \psi)$-worlds are relatively closer to w than $(\varphi \wedge \neg\psi)$-worlds. The case (A) means that the antecedent φ is *not* possible at w, i.e., $\Diamond\varphi$ is not true at w. In such a case, we make $\varphi \boxdot\!\!\rightarrow \psi$ *vacuously* true.

Remark 2.2. If $(W, \$, V)$ represents the comparative similarity of a given *agent*, we can also read $\varphi \boxdot\!\!\rightarrow \psi$ as 'conditional on φ, the agent believes that ψ'. Let us rewrite $\varphi \boxdot\!\!\rightarrow \psi$ as $B^{\varphi}\psi$. Let us use the construction of Remark 2.1. Then, (van Benthem 2010, Ch.13) defines the truth condition of $B^{\varphi}\psi$ at w as follows: ψ is true at all the $\leq_w$-minimal elements in $R(w) \cap [\![\varphi]\!]$. If we assume that $(W, \$, V)$ satisfies the condition called *limit assumption* (a kind of minimality condition, see also (Sano 2009, Definition 2) and (Lewis 1973, Section 1.4)), we can reformulate the truth condition of $\varphi \boxdot\!\!\rightarrow \psi$ into the simpler form as done in (van Benthem 2010, Ch.13). However, this paper does not assume the limit assumption, because we prefer mathematical generality.

2.2 A Motivation for Hybridizing David Lewis' Counterfactual Logic

A motivation of Sano (2009) for a hybridization of David Lewis' counterfactual logic is concerned with an *egocentric* interpretation of systems of spheres, where

$\varphi \boxdot\!\!\rightarrow \psi$ is useful to analyze the *contextually definite description*. In this section, we first explain this egocentric interpretation and its merit. Second, we motivate the hybridization of David Lewis' counterfactual logic.

(Lewis 1973, sec.5.3) considers Arthur Prior's egocentric reading of sentences and proposed that his counterfactual connective expresses *contextually definite descriptions* (e.g., 'The pig is pregnant'), whose logical form is '*The x such that* φ *is such that* ψ'. To be more accurate, he uses the connective $\varphi \boxdot\!\!\Rightarrow \psi$ defined as $\Diamond\varphi \wedge (\varphi \boxdot\!\!\rightarrow \psi)$, whose truth condition corresponds exactly to the case (B) of the truth condition of $\varphi \boxdot\!\!\rightarrow \psi$. According to this egocentric reading, the truth of a sentence is relativized to a thing or an individual, so the truth of sentence φ at x means that the individual x has the property φ. Then, a system of spheres around x represents its *comparative salience*, i.e., x's degree of familiarity between things and individuals. In this case, $\bigcup \$_x$ represents the set of all the things that are salient to x. Following Lewis, let us call $\bigcup \$_x$ x's *ken*. Then, 'The pig is grunting', formalized as '$\mathsf{Pig} \boxdot\!\!\Rightarrow \mathsf{Grunting}$', is true at an individual x iff the grunting pig is more salient for x than the non-grunting pigs. Furthermore, Lewis' analysis allows us to handle a *sequence* of egocentric

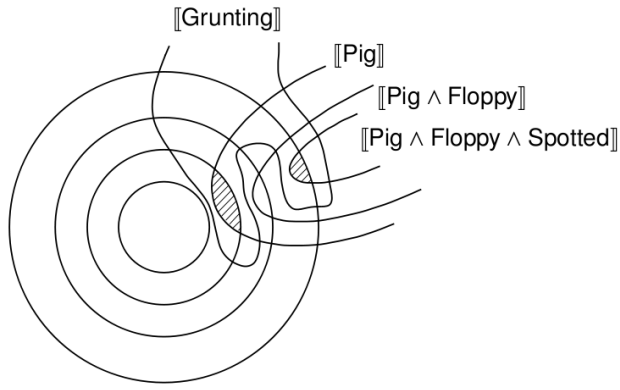


Figure 2: A Sequence of Egocentric Conditionals

conditionals (Lewis 1973, p.114): Suppose that you are walking past a piggery.

$$\begin{array}{c}\text{The pig is grunting: } \mathsf{Pig} \boxdot\!\!\Rightarrow \mathsf{Grunting},\\ \text{The pig with floppy ears is not grunting: } (\mathsf{Pig} \wedge \mathsf{Floppy}) \boxdot\!\!\Rightarrow \neg\mathsf{Grunting}, \text{ and}\\ \text{The spotted pig with floppy ears is grunting: } (\mathsf{Pig} \wedge \mathsf{Floppy} \wedge \mathsf{Spotted}) \boxdot\!\!\Rightarrow \mathsf{Grunting},\end{array} \tag{1}$$

where we use Pig, Grunting, etc. as propositional variables (see also Figure 2). According to the usual analysis of a definite description like $\mathsf{Grunting}(\imath x\mathsf{Pig}(x))$ (meaning that there exists a unique x such that $\mathsf{Pig}(x)$ and $\mathsf{Grunting}(x)$, where we use Pig and Grunting as unary predicate symbols in first-order logic), however, we cannot deal with such a sequence, since we can never make both $\mathsf{Grunting}(\imath x\mathsf{Pig}(x))$ and $\neg\mathsf{Grunting}(\imath x(\mathsf{Pig}(x) \wedge \mathsf{Floppy}(x)))$ true at the same time.

A hybridization of Lewis' counterfactual logic deserves investigation, because this hybridization enables us to regard the following inference as a process of revising local proposition (which depends on the given situation) by using global proposition (independent of the situation):

$$\frac{\begin{array}{c}\text{The pig is Mary.}\\ \text{Mary is pregnant.}\end{array}}{\text{Therefore: The pig is pregnant.}} \tag{2}$$

In this example, we regard the sentences containing 'the' as representing a local proposition, since the truth of these sentences depends on who utters them. On the other hand, we regard 'Mary is pregnant' as a global proposition, because its truth is independent of the speaker. To deal with the sentences containing 'the', we make use of Lewis' egocentric reading of these sentences mentioned above. In order to deal with sentences like the second premise , we need the modern hybrid formalism, whose roots can be traced back to Arthur Prior (see, e.g., Blackburn (2006)).

Hybrid systems introduce *nominals* i (names for states) and *satisfaction operators* $@_i p$ (p is true at the state named by i) and formalize 'Mary is pregnant' as $@_{\mathsf{MARY}}\mathsf{Pregnant}$. In reformulating Prior's egocentric reading, Lewis himself also deals with a similar kind of sentence (Lewis 1973, p.112):

> [...] the egocentric sentence 'x is such that (the Anighito meteorite is an x such that x is a rock)' is true at me, or anything else, because the Anighito meteorite is a rock;

Familiarity with hybrid formalism would allow Lewis to write this sentence in the most compact way possible: $@_{\mathsf{ANIGHITO\ METEORITE}}\mathsf{Rock}$. This explains the subtitle of Sano (2009): David Lewis Meets Arthur Prior Again.

2.3 A Complete Hilbert-style Axiomatization for Hybrid Counterfactual Logic

Let us introduce our syntax $\mathcal{HC}(@)$ for hybrid counterfactual logic. Let PROP be the set of all propositional variables, NOM the set of all *nominal variables*,

where we assume that $\mathsf{PROP} \cap \mathsf{NOM} = \emptyset$. Then the set $\mathsf{FORM}_{\mathcal{HC}(@)}$ of all formulas of $\mathcal{HC}(@)$ is defined inductively as:

$$\varphi ::= p \mid i \mid \neg\varphi \mid \varphi \wedge \psi \mid @_i\varphi \mid \varphi \boxdot\!\!\rightarrow \psi.$$

We denote by FORM_C the set of all non-hybrid formulas, i.e., all the formulas defined inductively from $\mathsf{PROP} \cup \{\neg, \wedge, \boxdot\!\!\rightarrow\}$. We also use $\bot$, $\top$, $\rightarrow$ and $\vee$ as the usual abbreviation and define $\Diamond\varphi$ as $\neg(\varphi \boxdot\!\!\rightarrow \bot)$ and $\varphi \boxdot\!\!\Rightarrow \psi$ as $\Diamond\varphi \wedge (\varphi \boxdot\!\!\rightarrow \psi)$. Over sphere models, we define the semantics for this hybrid counterfactual formalism as follows. A *sphere model* $\mathfrak{M} = (W, \$, V)$ consists of a system of spheres $(W, \$)$ and a *hybrid valuation* $V : \mathsf{PROP} \cup \mathsf{NOM} \to \mathcal{P}(W)$ satisfying $\#V(i) = 1$ for any $i \in \mathsf{Nom}$ (we usually write $V(i) = \{i^V\}$). Given a sphere model $\mathfrak{M} = (W, \$, V)$, we define $\mathfrak{M}, w \Vdash \varphi$ as follows:

$$\begin{array}{lll}
\mathfrak{M}, w \Vdash p & \text{iff} & w \in V(p) \\
\mathfrak{M}, w \Vdash i & \text{iff} & w = i^V \\
\mathfrak{M}, w \Vdash \neg\varphi & \text{iff} & \mathfrak{M}, w \nVdash \varphi \\
\mathfrak{M}, w \Vdash \varphi \wedge \psi & \text{iff} & \mathfrak{M}, w \Vdash \varphi \text{ and } \mathfrak{M}, w \Vdash \psi \\
\mathfrak{M}, w \Vdash @_i\varphi & \text{iff} & \mathfrak{M}, i^V \Vdash \varphi. \\
\mathfrak{M}, w \Vdash \varphi \boxdot\!\!\rightarrow \psi & \text{iff} & \begin{cases} \text{(A) } \bigcup \$_w \cap [\![\varphi]\!]_{\mathfrak{M}} = \emptyset \text{ or} \\ \text{(B) } (\exists S \in \$_w)\,(S \cap [\![\varphi]\!]_{\mathfrak{M}} \neq \emptyset \text{ and } S \cap [\![\varphi]\!]_{\mathfrak{M}} \subseteq [\![\psi]\!]_{\mathfrak{M}}), \end{cases}
\end{array}$$

where $[\![\varphi]\!]_{\mathfrak{M}} = \{w \in W \mid \mathfrak{M}, w \Vdash \varphi\}$ (we usually drop the subscript $\mathfrak{M}$ from $[\![\varphi]\!]_{\mathfrak{M}}$ when it is clear from the context). A *pointed sphere model* is a pair of a sphere model and a state from the domain of it. Given two pointed sphere model $\mathfrak{M}, w$ and $\mathfrak{M}, v$, we define $\mathfrak{M}, w \leftrightsquigarrow \mathfrak{M}, v$ (or, $\mathfrak{M}, w \leftrightsquigarrow_h \mathfrak{M}, v$) if $\mathfrak{M}, w \Vdash \varphi$ iff $\mathfrak{N}, v \Vdash \varphi$ for all $\varphi \in \mathsf{FORM}_C$ (or, all $\varphi \in \mathsf{FORM}_{\mathcal{HC}(@)}$, respectively). We say that φ is *valid* on $(W, \$)$ if $[\![\varphi]\!]_{(W,\$,V)} = W$ for any hybrid valuation V.

Let us return to our motivating example (2). We can formalize our motivating inference (2) as follows:

$$((\mathsf{Pig} \boxdot\!\!\Rightarrow \mathsf{MARY}) \wedge @_{\mathsf{MARY}}\mathsf{Pregnant}) \rightarrow (\mathsf{Pig} \boxdot\!\!\Rightarrow \mathsf{Pregnant}), \tag{3}$$

where we regard Pig and Pregnant as propositional variables and MARY as a nominal variable. In fact, our motivating example (3) is valid. For details, see (Sano 2009, p.521).

We can also give a Hilbert-style axiomatization $\mathbf{V}_{\mathcal{HC}(@)}$ in Table 1 to all the valid formulas on all systems of spheres. Our axiomatization extends Lewis' axiomatization **V** (Lewis 1973, ch.6), which consists of the axioms: **CT**, **ID**, **MOD**, **ARR** and the inference rule: **MP**, **DwC**, **ILE**, and the uniform substitutions.

Axioms for $\mathbf{V}_{\mathcal{HC}(@)}$	
CT	$\vdash \varphi$, for all classical tautologies φ
K@	$\vdash @_i(p \to q) \to (@_i p \to @_i q)$
Self-Dual	$\vdash \neg @_i p \leftrightarrow @_i \neg p$
Ref	$\vdash @_i i$
Intro	$\vdash (i \wedge p) \to @_i p$
Agree	$\vdash @_i @_j p \to @_j p$
Back	$\vdash @_i p \to (q \boxdot\!\!\to @_i p)$
ID	$\vdash p \boxdot\!\!\to p$
MOD	$\vdash (\neg p \boxdot\!\!\to p) \to (q \boxdot\!\!\to p)$
ARR	$\vdash \neg(p \boxdot\!\!\to \neg q) \to [((p \wedge q) \boxdot\!\!\to r) \leftrightarrow (p \boxdot\!\!\to (q \to r))]$
Rules for $\mathbf{V}_{\mathcal{HC}(@)}$	
MP	If $\vdash \varphi \to \psi$ and $\vdash \varphi$, then $\vdash \psi$
DwC	If $\vdash (\theta_1 \wedge \cdots \wedge \theta_n) \to \psi$, then $\vdash ((\varphi \boxdot\!\!\to \theta_1) \wedge \cdots \wedge (\varphi \boxdot\!\!\to \theta_n)) \to (\varphi \boxdot\!\!\to \psi)$ $(n \geq 1)$
ILE	If $\vdash \varphi \leftrightarrow \psi$, then $\vdash (\varphi \boxdot\!\!\to \theta) \leftrightarrow (\psi \boxdot\!\!\to \theta)$.
Nec@	If $\vdash \varphi$, then $\vdash @_i \varphi$
Sub	If $\vdash \varphi$, then $\vdash \sigma(\varphi)$, where σ denotes a substitution that uniformly replaces propositional variables by formulas and nominals by nominals.

Table 1: The Axiomatization $\mathbf{V}_{\mathcal{HC}(@)}$

While Lewis (1973) shows that **V** enjoys soundness and completeness with respect to all finite sphere models (hence decidability), Sano (2009) established:

Theorem 1. *$\mathbf{V}_{\mathcal{HC}(@)}$ is sound and complete with respect to all finite sphere models, i.e., for any $\varphi \in \mathsf{FORM}_{\mathcal{HC}(@)}$, φ is valid on all finite sphere models iff φ is a theorem of $\mathbf{V}_{\mathcal{HC}(@)}$. Therefore, $\mathbf{V}_{\mathcal{HC}(@)}$ is decidable.*

Proof. See (Sano 2009, Proposition 1 and Theorem 2). □

2.4 Sphere-Bisimulation

We can also define the notion of *bisimulation* between sphere-models (cf. the notion of *topo-bisimulation* (Blackburn and van Benthem 2007, p.77, Definition 53)). Unlike (Sano 2009, Definition 12), we also define the appropriate notion of bisimulation for hybrid counterfactual syntax here. The reader can find the notion of bisimulation for hybrid languages in (ten Cate 2005, Definition 4.1.1).

Definition 2.2. Let $\mathfrak{M} = (W, \$, V)$ and $\mathfrak{N} = (W', \$', V')$ be sphere models. $Z \subseteq W \times W'$ is *a sphere-bisimulation* between $\mathfrak{M}$ and $\mathfrak{N}$ if, for any $(w, w') \in W \times W'$, wZw' implies:

(prop) $[w \in V(p)$ iff $w' \in V'(p)]$ for any $p \in \mathsf{PROP}$.

(forth) If $S \in \$_w$ and $S \cap X \neq \emptyset$, then $(\exists\, S' \in \$_{w'})\,[S' \subseteq Z[S]$ and $S' \cap Z[X] \neq \emptyset]$ $(X \subseteq W)$.

(back) If $S' \in \$'_{w'}$ and $S' \cap Y \neq \emptyset$, then $(\exists\, S \in \$_w)\,[S \subseteq Z^{-1}[S']$ and $S \cap Z^{-1}[Y] \neq \emptyset]$ $(Y \subseteq W')$.

where $Z[X] = \{\, y \in W' \mid (\exists\, x \in X)\,(xZy) \,\}$ and $Z^{-1}[Y] = \{\, x \in W \mid (\exists\, y \in Y)\,(xZy) \,\}$. We say that $Z \subseteq W \times W'$ is a *hybrid sphere-bisimulation* between $\mathfrak{M}$ and $\mathfrak{N}$ if Z is a sphere-bisimulation and it also satisfies the following:

(nom) wZw' implies $(w \in V(i)$ iff $w' \in V'(i))$ for any $i \in \mathsf{NOM}$.

(sat) $i^V Z i^{V'}$ for any $i \in \mathsf{NOM}$.

Let $\mathfrak{M}, w$ and $\mathfrak{N}, v$ be pointed sphere models. We say that $\mathfrak{M}, w$ and $\mathfrak{N}, v$ are *sphere-bisimilar* (notation: $\mathfrak{M}, w \leftrightarrows \mathfrak{N}, v$) if there exists a sphere-bisimulation Z such that wZv. Similarly, we define the notion of *hybrid sphere-bisimilarity* between $\mathfrak{M}, w$ and $\mathfrak{N}, v$ and denote it by $\mathfrak{M}, w \leftrightarrows_h \mathfrak{N}, v$.

Remark 2.3. Demey (2010) initiated a systematic exploration of the model theory of epistemic plausibility models. He defined various notions of bisimulations parameterized by a language, but his notion (Demey 2010, Definition 8) of bisimulation for epistemic plausibility models assumes the minimality assumption (recall Remark 2.2). Our notion of sphere-bisimulation, however, does not depend on such an assumption, i.e., the Limit assumption, though we assume the condition (S3) of sphere systems which amounts to connectedness of $\leq_w$ (see Remark 2.1). It would be interesting to reformulate our notion of sphere-bisimulation in terms of epistemic plausibility models and see if the connectedness of $\leq_w$ is essential.

Example 1. (a) This example is taken from (Sano 2009, Example 2). Let us consider the following sphere models $(W, \$, V)$ and $(W', \$', V')$: $W = \{a, b\}$, $\$_a = \{\emptyset, \{b\}\}$, $\$_b = \{\emptyset, \{a\}\}$, $V(p) = \{a, b\}$ $(p \in \mathsf{Prop})$; $W' = \{0\}$, $\$'_0 = \{\emptyset, \{0\}\}$, $V'(p) = \{0\}$ $(p \in \mathsf{Prop})$. Define $Z \subseteq W \times W' := \{(a, 0), (b, 0)\}$. Then, Z is a sphere-bisimulation between $(W, \$, V)$ and $(W', \$', V')$.

(b) Let us also give an example to the notion of hybrid sphere-bisimulation. Define $(W, \$, V)$ and $(W', \$', V')$ as follows: $W = \{a, b, c\}$, $\$_a = \{\emptyset, \{b\}, \{b, c\}\}$, $\$_b = \{\emptyset, \{a\}, \{a, c\}\}$, $\$_c = \{\emptyset\}$, $V(p) = \{a, b\}$ $(p \in \mathsf{PROP})$, $V(i) = \{c\}$ $(i \in \mathsf{NOM})$; $W' = \{0, 1\}$, $\$'_0 = \{\emptyset, \{0\}\}$, $\$'_1 = \{\emptyset\}$, $V'(p) = \{0\}$ $(p \in \mathsf{PROP})$, $V'(i) = \{1\}$

($i \in \mathsf{NOM}$). Define $Z \subseteq W \times W' := \{(a, 0), (b, 0), (c, 1)\}$. Then, Z is a hybrid sphere-bisimulation between $(W, \$, V)$ and $(W', \$', V')$.

Proposition 1. *Let* $\mathfrak{M} = (W, \$, V)$ *and* $\mathfrak{N} = (W', \$', V')$ *be sphere models.*

(i) *If* Z *is a sphere-bisimulation between* $\mathfrak{M}$ *and* $\mathfrak{N}$*, then* wZv *implies* $\mathfrak{M}, w \leftrightsquigarrow \mathfrak{N}, v$ *for all* $(w, v) \in W \times W'$.

(ii) *If* Z *is a hybrid sphere-bisimulation between* $\mathfrak{M}$ *and* $\mathfrak{N}$*, then* wZv *implies* $\mathfrak{M}, w \leftrightsquigarrow_h \mathfrak{N}, v$ *for all* $(w, v) \in W \times W'$.

Proof. (i) is the same as (Sano 2009, p.535, Proposition 5). So, the proof of (i) can be found there. Let us establish (ii). By induction on φ, we show that wZv implies $\mathfrak{M}, w \Vdash \varphi$ iff $\mathfrak{N}, v \Vdash \varphi$ for all $(w, v) \in W \times W'$. It suffices to check the case $\varphi \equiv @_i\psi$ (we can establish the case where $\varphi \equiv i$ by (nom)). Suppose wZv. Then, $\mathfrak{M}, w \Vdash @_i\psi$ iff $\mathfrak{M}, i^V \Vdash \psi$ iff $\mathfrak{N}, i^{V'} \Vdash \psi$ by induction hypothesis and (sat) iff $\mathfrak{N}, v \Vdash @_i\psi$. □

3 Dynamified Hybrid Counterfactual Logic

3.1 A Motivation for Dynamifying David Lewis' Counterfactual Logic

Lewis states that comparative salience is easy to change as follows:

> I am speaking of how salient these things were *before* I started to think up examples of things that were not very salient; comparative salience is much shiftier even than comparative similarity. (Lewis 1973, p.113)

Lewis does not explain how we can obtain a new comparative salience from an old one in the formal setting. However, a dynamification of (hybrid) counterfactual logic enables us to describe *how* $\$_w$ of comparative salience changes. In what follows in this section, we first see why Lewis requires that comparative salience is shifty in connection with the contextually definite description. Second, we explain another motivation for the dynamification, which will provide us with a desirable explanation of a sequence of egocentric conditionals (1) (recall also Figure 2). Finally, we introduce our syntax for a dynamification of hybrid counterfactual logic.

As we mentioned above, Lewis regards $\mathsf{Pig} \mathrel{\Box\!\!\Rightarrow} \mathsf{Grunting}$ as the formalization of 'the x such that Pig *is* such that $\mathsf{Grunting}$', i.e., 'the pig *is* grunting'. It is

natural to understand that the sentence 'the pig *is* grunting' presupposes that there is a unique most salient pig to x (Lewis 1973, p.116). However, we can consider the case where there are *two equally most salient* pigs and both of them are grunting, hence $\mathsf{Pig} \boxdot\!\!\Rightarrow \mathsf{Grunting}$ is true at x. In such a case, we would face a presupposition failure. According to Lewis, however, we almost never have a tie, since comparative salience is quite shifty as follows:

> Consider that comparative salience is shifty in the extreme. Nothing is easier than to break the tie; and if it were broken *either way* the sentence would be true. Recognizing the inevitable vagueness of comparative salience, we see that we almost never will simply have a tie. (Lewis 1973, p.116)

In this sense, it is important for Lewis to give an illustration to changes of comparative salience. Modern developments of dynamic epistemic logic (cf. van Ditmarsch et al. (2008), van Benthem (2010)) allow us to handle several kinds of comparative salience change by dynamic modalities (we will later show concrete examples in Example 2 and Example 4).

Another motivation for a dynamification of Lewis' counterfactual logic is concerned with a sequence of egocentric conditionals. It is easy to see that such a sequence involves some change of the speaker's attention. For example, if the speaker first says 'the pig is grunting' and then says 'the pig with floppy ears is not grunting', then it is natural to regard that he or she changes his or her attention from one pig to another pig. As we will see in the next section (see Example 3), an idea from dynamic epistemic logic gives us a logical framework to capture such attention change.

As for dynamic modalities, this paper considers $[\varphi!]$ and $[\Uparrow \varphi]$ in our sphere-model setting, which amount to the update using link-cutting (van Benthem and Liu 2007, Yamada 2006) and the radical upgrade (van Benthem 2010, Section 15.7), respectively. Let us define the set $\mathsf{FORM}_{\mathcal{DHC}(@)}$ of all formulas of dynamified hybrid counterfactual logic by the mutual induction as follows:

$$\varphi ::= p \mid i \mid \neg\varphi \mid \varphi \wedge \psi \mid @_i\varphi \mid \varphi \boxdot\!\!\rightarrow \psi \mid [\pi]\varphi,$$
$$\pi ::= \varphi! \mid \Uparrow \varphi.$$

We denote by $\mathsf{FORM}_{\mathcal{DC}}$ the set of all *non-hybrid* formulas, i.e., all formulas not containing any hybrid vocabulary: both nominals and satisfaction operators. In what follows, we will explain the truth conditions of $[\alpha!]\varphi$ and $[\Uparrow \alpha]\varphi$ in terms of sphere-models.

3.2 Salience Update in Sphere Models

Definition 3.1 (Salience Update)**.** Let $\mathfrak{M} = (W, \$, V)$ and $\alpha \in \mathsf{FORM}_{\mathcal{DHC}(@)}$. We define:

$$\mathfrak{M}, w \Vdash [\alpha!]\varphi \quad \text{iff} \quad \mathfrak{M}^{\alpha!}, w \Vdash \varphi,$$

where the *salience update* $\mathfrak{M}^{\alpha!} = (W^{\alpha!}, \$^{\alpha!}, V^{\alpha!})$ is defined as:

- $W^{\alpha!} := W$.
- $\$^{\alpha!} : W \to \mathcal{PP}(W)$ is defined by: for any $w \in W$, $\$^{\alpha!}_w := \{ S \cap [\![\alpha]\!]_{\mathfrak{M}} \mid S \in \$_w \}$.
- $V^{\alpha!}(a) := V(a)$ for any $a \in \mathsf{PROP} \cup \mathsf{NOM}$.

Proposition 2. $(W^{\alpha!}, \$^{\alpha!})$ *is a system of spheres.*

In the above definition, the domain $W^{\alpha!}$ is not $[\![\alpha]\!]$ but the original domain W. So, this definition '*cuts*' the system $\$_w$ of spheres similarly to the update using link-cutting in (van Benthem and Liu 2007) and (Yamada 2006). Mathematically speaking, our truth condition of $[\alpha!]\varphi$ is similar to the one given in (Yamada 2006), while (van Benthem and Liu 2007) requires the assumption $\mathfrak{M}, w \Vdash \alpha$ in the truth condition of $[\alpha!]\varphi$ (see also (van Benthem 2010, Section 15.6)) [2].

Compared to the ordinary definition of public announcement update by removing all the $\neg\alpha$-states, a link-cutting technique has a conceptual merit for handling a sequence of egocentric conditionals (Example 3 below). Moreover, it has a merit over hybrid language. If we remove all the $\neg\alpha$-states and the denotation of i is a $\neg\alpha$-state, the denotation of i becomes empty in the updated model, which implies that the updated valuation function does not satisfy the requirement for *hybrid valuation*: a nominal i should be true at *exactly one* state. In order to handle such situation, we need to rebuild the static hybrid language allowing the possibility $\#V(i) \leq 1$ (for such a study, the reader can refer to Ulrik Hansen (2011)). However, link-cutting keep the original domain, and so, we can avoid the technical difficulty about the denotation of nominals.

Example 2. Figure 3 gives an example of salience update. In the first figure of Figure 3, there are two equally most salient pigs and both of them are grunting to the speaker w. Suppose that the upper pig is Mary and the lower pig is Sally. The circle around Sally represents the denotation of Black, and so, we assume that the unique black thing is Sally in this model. Salience update by [Black!]

[2] Independently of (Yamada 2006, van Benthem and Liu 2007) in the DEL framework, Lewis himself also proposed the link-cutting update having the same underlying idea as (Yamada 2006), when he considered the meaning of commanding (Lewis 1979, p.167).

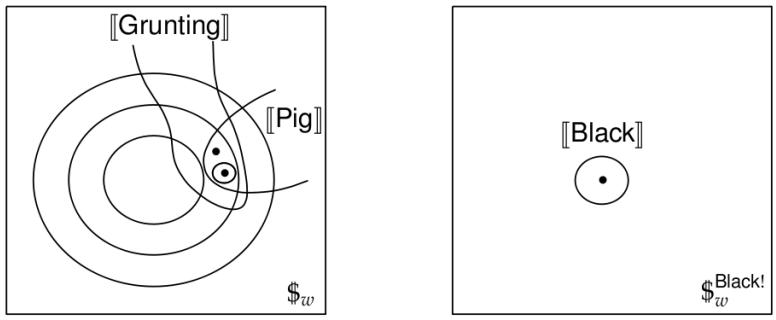


Figure 3: How to break tie by [Black!]

breaks the tie between Mary and Sally and makes Pig $\Box\!\!\Rightarrow$ Grunting (i.e., 'the pig is grunting') true after the update. This explains why [Black!](Pig $\Box\!\!\Rightarrow$ Grunting) is true at w on the first figure.

In the first figure, we can say that 'the non-black pig is grunting' is true, since the most salient non-black pig is Mary alone. However, after the update by [Black!], 'the non-black pig is grunting' is no longer true, because all the non-black pigs are outside of w' ken, i.e., $\$_w^{\mathsf{Black!}} = \llbracket \mathsf{SALLY} \rrbracket$. This can be summarized as: $(\neg \mathsf{Black} \wedge \mathsf{Pig}) \Box\!\!\Rightarrow \mathsf{Grunting}$ is true at w but $[\mathsf{Black!}]((\neg \mathsf{Black} \wedge \mathsf{Pig}) \Box\!\!\Rightarrow \mathsf{Grunting})$ is false at w in the first figure.

$[\alpha!]p$	$\leftrightarrow$	p
$[\alpha!]i$	$\leftrightarrow$	i
$[\alpha!]\neg\varphi$	$\leftrightarrow$	$\neg[\alpha!]\varphi$
$[\alpha!](\varphi \wedge \psi)$	$\leftrightarrow$	$[\alpha!]\varphi \wedge [\alpha!]\psi$
$[\alpha!]@_i\varphi$	$\leftrightarrow$	$@_i[\alpha!]\varphi$
$[\alpha!](\varphi \Box\!\!\rightarrow \psi)$	$\leftrightarrow$	$(\alpha \wedge [\alpha!]\varphi) \Box\!\!\rightarrow [\alpha!]\psi$

Table 2: Reduction Axioms for $[\alpha!]$

Remark 3.1. In the sixth axiom of Table 2, let us rewrite $\varphi \Box\!\!\rightarrow \psi$ as the relative belief operator $B^{\varphi}\psi$ (recall Remark 2.2). Then, we obtain $[\alpha!]B^{\varphi}\psi \leftrightarrow B^{\alpha \wedge [\alpha!]\varphi}[\alpha!]\psi$.

As for the public announcement update with eliminating all the $\neg\alpha$-states, the reader can find the reduction axiom, e.g., in (van Benthem 2010, Section 15.6). In our notation, we can write it as: $[\alpha!]B^{\varphi}\psi \leftrightarrow (\alpha \to B^{\alpha\wedge[\alpha!]\varphi}[\alpha!]\psi)$. Since we do not require the assumption $\mathfrak{M}, w \Vdash \alpha$ in the truth condition of $[\alpha!]\varphi$, our reduction axiom does not have the assumption α.

Proposition 3. *All the axioms in Table 2 are valid.*

Proof. We check the validity of the fifth and sixth axioms alone. First, we can show the validity of the fifth axiom as follows: $\mathfrak{M}, w \Vdash [\alpha!]@_i\varphi$ iff $\mathfrak{M}^{\alpha!}, w \Vdash @_i\varphi$ iff $V^{\alpha!}(i) \subseteq [\![\varphi]\!]_{\mathfrak{M}^{\alpha!}}$ iff $V(i) \subseteq [\![[\alpha!]\varphi]\!]_{\mathfrak{M}}$ (by the definition of $V^{\alpha!}$ and the truth condition of $[\alpha!]\varphi$) iff $\mathfrak{M}, w \Vdash @_i[\alpha!]\varphi$.

Second, let us establish the validity of the sixth axiom. We can proceed as follows: $\mathfrak{M}, w \Vdash [\alpha!](\varphi \boxdot\!\!\to \psi)$ iff $\mathfrak{M}^{\alpha!}, w \Vdash \varphi \boxdot\!\!\to \psi$ iff:

$$\begin{cases} \bigcup \$_w^{\alpha!} \cap [\![\varphi]\!]_{\mathfrak{M}^{\alpha!}} = \emptyset \text{ or} \\ (\exists S \in \$_w)\,[S \cap [\![\alpha]\!]_{\mathfrak{M}} \cap [\![\varphi]\!]_{\mathfrak{M}^{\alpha!}} \neq \emptyset \text{ and } S \cap [\![\alpha]\!]_{\mathfrak{M}} \cap [\![\varphi]\!]_{\mathfrak{M}^{[\alpha!]}} \subseteq [\![\psi]\!]_{\mathfrak{M}^{\alpha!}}]. \end{cases}$$

Since $\bigcup \$_w \cap [\![\alpha]\!]_{\mathfrak{M}} = \bigcup \$_w^{\alpha!}$ and $[\![\gamma]\!]_{\mathfrak{M}^{\alpha!}} = [\![[\alpha!]\gamma]\!]_{\mathfrak{M}}$ for any γ, the above disjunction is equivalent to:

$$\begin{cases} \bigcup \$_w \cap [\![\alpha]\!]_{\mathfrak{M}} \cap [\![[\alpha!]\varphi]\!]_{\mathfrak{M}} = \emptyset \text{ or} \\ (\exists S \in \$_w)\,[S \cap [\![\alpha]\!]_{\mathfrak{M}} \cap [\![[\alpha!]\varphi]\!]_{\mathfrak{M}} \neq \emptyset \text{ and } S \cap [\![\alpha]\!]_{\mathfrak{M}} \cap [\![[\alpha!]\varphi]\!]_{\mathfrak{M}} \subseteq [\![[\alpha!]\psi]\!]_{\mathfrak{M}}]. \end{cases}$$

Equivalently, we obtain $\mathfrak{M}, w \Vdash (\alpha \wedge [\alpha!]\varphi) \boxdot\!\!\to [\alpha!]\psi$, as required. □

Example 3. Let us go back to a sequence of egocentric conditionals (1) (see also Figure 2). By Proposition 3, it is easy to verify, e.g., that $(\mathsf{Pig} \wedge \mathsf{Floppy}) \boxdot\!\!\Rightarrow \neg\mathsf{Grunting}$ is equivalent to $[\mathsf{Floppy!}](\mathsf{Pig} \boxdot\!\!\Rightarrow \neg\mathsf{Grunting})$ (recall that $\varphi \boxdot\!\!\Rightarrow \psi \equiv \Diamond\varphi \wedge (\varphi \boxdot\!\!\to \psi)$ and $\Diamond\varphi \equiv \neg(\varphi \boxdot\!\!\to \bot)$). We have explained that generating such a sequence involves the speaker's attention change. By the equivalence above, we can regard the updated comparative salience $\$_w^{\alpha!}$ as a way of expressing the result of the speaker w's attention change. Then, we can rewrite our sequence as follows [3]:

The pig is grunting: $\mathsf{Pig} \boxdot\!\!\Rightarrow \mathsf{Grunting}$,
The pig with floppy ears is not grunting: $[\mathsf{Floppy!}](\mathsf{Pig} \boxdot\!\!\Rightarrow \neg\mathsf{Grunting})$, and
The spotted pig with floppy ears is grunting: $[\mathsf{Spotted!}][\mathsf{Floppy!}](\mathsf{Pig} \boxdot\!\!\Rightarrow \neg\mathsf{Grunting})$,

[3] Remark that $[\mathsf{Spotted!}][\mathsf{Floppy!}](p \boxdot\!\!\Rightarrow q)$ is equivalent to $[(\mathsf{Floppy} \wedge \mathsf{Spotted})!](p \boxdot\!\!\Rightarrow q)$ hence $[\mathsf{Floppy!}][\mathsf{Spotted!}](p \boxdot\!\!\Rightarrow q)$.

In this way, iteration of salience update operators provide a way of capturing the result of the speaker's attention change.

Similarly to public announcement update (van Benthem 2010, Section 15.5), our salience update *respects* sphere-bisimulation as follows.

Proposition 4. *Let $\mathfrak{M}, w$ and $\mathfrak{N}, v$ be pointed sphere models. Given any $\alpha \in \mathsf{FORM}_{\mathcal{DC}}$, if $\mathfrak{M}, w \leftrightarrows \mathfrak{N}, v$, then $\mathfrak{M}^{\alpha!}, w \leftrightarrows \mathfrak{N}^{\alpha!}, v$. Moreover, given any $\alpha \in \mathsf{FORM}_{\mathcal{DHC}(@)}$, if $\mathfrak{M}, w \leftrightarrows_h \mathfrak{N}, v$, then $\mathfrak{M}^{\alpha!}, w \leftrightarrows_h \mathfrak{N}^{\alpha!}, v$.*

Proof. It suffices to show the first part. Let Z be a sphere-bisimulation between $\mathfrak{M}$ and $\mathfrak{N}$ such that wZv. We show that it is also a sphere-bisimulation between $\mathfrak{M}^{\alpha!}$ and $\mathfrak{N}^{\alpha!}$. We need to check three conditions (prop), (forth), and (back) in Definition 2.2. Here let us focus on (forth), since (prop) is easy to show and we can establish (back) similarly to (forth).

Assume that wZv and that $S \in \$_w$ and $S \cap [\![\alpha]\!]_{\mathfrak{M}} \cap X \neq \emptyset$, where $X \subseteq W$. Our goal is to establish that there exists some $S' \in \$'_v$ such that (i) $S' \cap [\![\alpha]\!]_{\mathfrak{N}} \subseteq Z[S \cap [\![\alpha]\!]_{\mathfrak{M}}]$ and (ii) $S' \cap [\![\alpha]\!]_{\mathfrak{N}} \cap Z[X] \neq \emptyset$. Since Z is a sphere-bisimulation between $\mathfrak{M}$ and $\mathfrak{N}$ such that wZv, we can choose $S' \in \$'_v$ such that $S' \subseteq Z[S]$ and $S' \cap Z[[\![\alpha]\!]_{\mathfrak{M}} \cap X] \neq \emptyset$. This is our witness.

First, we establish (i). Fix any $y \in S' \cap [\![\alpha]\!]_{\mathfrak{N}}$. Since $S' \subseteq Z[S]$, $y \in Z[S]$, i.e., xZy and $x \in S$ for some $x \in W$. Fix such x. We want to show $x \in [\![\alpha]\!]_{\mathfrak{M}}$, because this gives us $y \in Z[S \cap [\![\alpha]\!]_{\mathfrak{M}}]$. By Proposition 1 (i) [4], we deduce from xZy and $y \in [\![\alpha]\!]_{\mathfrak{N}}$ that $x \in [\![\alpha]\!]_{\mathfrak{M}}$, as desired.

Second, let us demonstrate (ii). We have obtained $S' \cap Z[[\![\alpha]\!]_{\mathfrak{M}} \cap X] \neq \emptyset$. So, choose some $y \in S' \cap Z[[\![\alpha]\!]_{\mathfrak{M}} \cap X]$. It is easy to see that $y \in S' \cap Z[X]$. So, it suffices to show $y \in [\![\alpha]\!]_{\mathfrak{N}}$. $y \in Z[[\![\alpha]\!]_{\mathfrak{M}} \cap X]$ implies that $x \in [\![\alpha]\!]_{\mathfrak{M}} \cap X$ and xZy for some $x \in W$. By Proposition 1 (i), we deduce from xZy and $x \in [\![\alpha]\!]_{\mathfrak{M}}$ that $y \in [\![\alpha]\!]_{\mathfrak{N}}$. □

3.3 Salience Upgrade in Sphere Models

Definition 3.2 (Radical Upgrade)**.** Let $\mathfrak{M} = (W, \$, V)$ and $\alpha \in \mathsf{FORM}_{\mathcal{DHC}(@)}$. We define:

$$\mathfrak{M}, w \Vdash [\Uparrow \alpha]\varphi \quad \text{iff} \quad \mathfrak{M}^{\Uparrow\alpha}, w \Vdash \varphi,$$

where the *radical upgrade* $\mathfrak{M}^{\Uparrow\alpha} = (W^{\Uparrow\alpha}, \$^{\Uparrow\alpha}, V^{\Uparrow\alpha})$ is defined as:

- $W^{\Uparrow\alpha} := W$.

[4] α might contain the salience update $[\beta!]$ but we can rewrite it to a formula of $\mathsf{FORM}_{\mathcal{C}}$ by Proposition 3. So, we can apply Proposition 1 (i) here.

- $\$^{\Uparrow\alpha} : W \to \mathcal{PP}(W)$ is defined by: for any $w \in W$,

$$\$^{\Uparrow\alpha}_w := \{ S \cap [\![\alpha]\!]_{\mathfrak{M}} \mid S \in \$_w \} \cup \left\{ \left(\bigcup \$_w \cap [\![\alpha]\!]_{\mathfrak{M}} \right) \cup S \mid S \in \$_w \right\}.$$

- $V^{\Uparrow\alpha}(a) := V(a)$ for any $a \in \mathsf{PROP} \cup \mathsf{NOM}$.

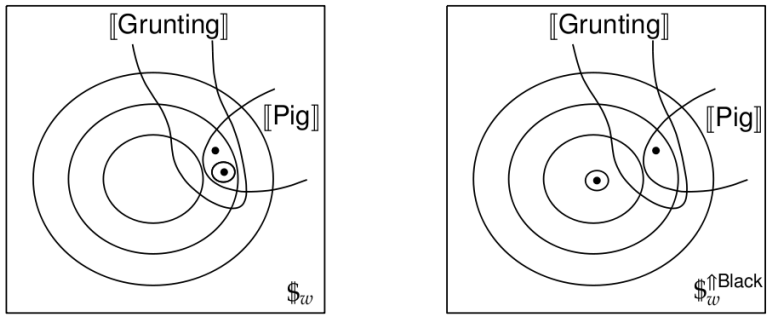


Figure 4: How to break tie by [⇑ Black]

Proposition 5. $(W^{\Uparrow\alpha}, \$^{\Uparrow\alpha})$ *is a system of spheres.*

Example 4. Figure 4 gives an example of radical upgrade. As for the first figure of Figure 4, we make the same assumption as in Example 2. The second figure of Figure 4 represents the radical upgrade by [⇑ Black]. All the black things (Sally alone, in this case) become more salient than all the non-black things, but we still keep all the non-black things of w's ken $\bigcup \$_w$ together with the previous comparative salience (similarity) structure even in our new $\bigcup \$^{\Uparrow\mathsf{Black}}_w$. Similarly to Example 2, the radical upgrade by [⇑ Black] breaks the tie between Mary and Sally and makes Pig $\Box\!\!\Rightarrow$ Grunting (i.e., 'the pig is grunting') true. This explains why [⇑ Black](Pig $\Box\!\!\Rightarrow$ Grunting) is true at w on the first figure.

Let us see one difference from the salience update by [Black!]. In the first figure of Example 2, both (¬Black ∧ Pig) $\Box\!\!\Rightarrow$ Grunting and ¬[Black!]((¬Black ∧ Pig) $\Box\!\!\Rightarrow$ Grunting) are true at w. After the *radical* upgrade by [⇑ Black], however, we keep the all the non-black things of w's ken $\bigcup \$_w$ in our new $\bigcup \$^{\Uparrow\mathsf{Black}}_w$, and so, 'the non-black pig is grunting' is *still true*. That is, [⇑ Black]((¬Black ∧ Pig) $\Box\!\!\Rightarrow$

Grunting) is true at w in the first figure. In this sense, the radical upgrade by $[\Uparrow$ Black$]$ does not change the truth of 'the non-black pig is grunting'.

$[\Uparrow \alpha]p$	$\leftrightarrow$	p
$[\Uparrow \alpha]i$	$\leftrightarrow$	i
$[\Uparrow \alpha]\neg\varphi$	$\leftrightarrow$	$\neg[\Uparrow \alpha]\varphi$
$[\Uparrow \alpha](\varphi \wedge \psi)$	$\leftrightarrow$	$[\Uparrow \alpha]\varphi \wedge [\Uparrow \alpha]\psi$
$[\Uparrow \alpha]@_i\varphi$	$\leftrightarrow$	$@_i[\Uparrow \alpha]\varphi$
$[\Uparrow \alpha](\varphi \mathrel{\Box\!\!\rightarrow} \psi)$	$\leftrightarrow$	$(\Diamond(\alpha \wedge [\Uparrow \alpha]\varphi) \wedge ((\alpha \wedge [\Uparrow \alpha]\varphi) \mathrel{\Box\!\!\rightarrow} [\Uparrow \alpha]\psi)) \vee$ $(\neg\Diamond(\alpha \wedge [\Uparrow \alpha]\varphi) \wedge ([\Uparrow \alpha]\varphi \mathrel{\Box\!\!\rightarrow} [\Uparrow \alpha]\psi))$

Table 3: Reduction Axioms for $[\Uparrow \alpha]$

Remark 3.2. Similarly to Remark 3.1, let us rewrite $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ as the relative belief operator $B^{\varphi}\psi$ in the sixth axiom of Table 3. Then, the resulting axiom is the same as the one given in (van Benthem 2010, Section 15.7).

Proposition 6. *All the axioms in Table 3 are valid.*

Proof. We check the validity of the sixth axiom alone (as for the fifth axiom, we can show the validity of it similarly to the proof of Proposition 3). Since $\bigcup \$_w = \bigcup \$_w^{\Uparrow\alpha}$ and $[\![[\Uparrow \alpha]\varphi]\!]_{\mathfrak{M}} = [\![\varphi]\!]_{\mathfrak{M}^{\Uparrow\alpha}}$, it is easy to see that $\mathfrak{M}, w \Vdash [\Uparrow \alpha](\varphi \mathrel{\Box\!\!\rightarrow} \psi)$ is equivalent to the disjunction of the following:

(A') $\bigcup \$_w \cap [\![[\Uparrow \alpha]\varphi]\!]_{\mathfrak{M}} = \emptyset$.

(B') $(\exists P \in \$_w^{\Uparrow\alpha})\,(P \cap [\![[\Uparrow \alpha]\varphi]\!]_{\mathfrak{M}} \neq \emptyset$ and $P \cap [\![[\Uparrow \alpha]\varphi]\!]_{\mathfrak{M}} \subseteq [\![[\Uparrow \alpha]\psi]\!]_{\mathfrak{M}})$.

Let us put

$$\gamma_1 \equiv \Diamond(\alpha \wedge [\Uparrow \alpha]\varphi) \wedge ((\alpha \wedge [\Uparrow \alpha]\varphi) \mathrel{\Box\!\!\rightarrow} [\Uparrow \alpha]\psi),$$
$$\gamma_2 \equiv \neg\Diamond(\alpha \wedge [\Uparrow \alpha]\varphi) \wedge ([\Uparrow \alpha]\varphi \mathrel{\Box\!\!\rightarrow} [\Uparrow \alpha]\psi).$$

Our goal is to show the equivalence between ((A') or (B')) and $\mathfrak{M}, w \Vdash \gamma_1 \vee \gamma_2$. In what follows in this proof, we drop the subscript from $[\![-]\!]_{\mathfrak{M}}$ and just write $[\![-]\!]$. Let us divide our argument into the following two cases: (i) $\mathfrak{M}, w \Vdash \Diamond(\alpha \wedge [\Uparrow \alpha]\varphi)$; (ii) $\mathfrak{M}, w \nVdash \Diamond(\alpha \wedge [\Uparrow \alpha]\varphi)$.

(i) Our assumption (i) is equivalent to $\bigcup \$_w \cap [\![\alpha \wedge [\Uparrow \alpha]\varphi]\!] \neq \emptyset$. Since it implies $\bigcup \$_w \cap [\![[\Uparrow \alpha]\varphi]\!] \neq \emptyset$, it suffices to establish the equivalence between (B')

and $\mathfrak{M}, w \Vdash \gamma_1$ for our desired equivalence above. By the definition of $\$_w^{\Uparrow\alpha}$, (B′) is equivalent to [5]:

$$(\exists S \in \$_w)(S \cap [\![\alpha \wedge [\Uparrow \alpha]\varphi]\!] \neq \emptyset \text{ and } S \cap [\![\alpha \wedge [\Uparrow \alpha]\varphi]\!] \subseteq [\![[\Uparrow \alpha]\psi]\!]).$$

Equivalently, $\mathfrak{M}, w \Vdash (\alpha \wedge [\Uparrow \alpha]\varphi) \boxdot\!\!\rightarrow [\Uparrow \alpha]\psi$. By our assumption (i), this is also equivalent to $\mathfrak{M}, w \Vdash \gamma_1$, as desired.

(ii) Our assumption is equivalent to $\bigcup \$_w \cap [\![\alpha \wedge [\Uparrow \alpha]\varphi]\!] = \emptyset$. In this case, we need to use another argument by cases: (A′) $\bigcup \$_w \cap [\![[\Uparrow \alpha]\varphi]\!] = \emptyset$; (∼ A′) $\bigcup \$_w \cap [\![[\Uparrow \alpha]\varphi]\!] \neq \emptyset$. First, consider the case (A′). In this case, we suffice to establish $\mathfrak{M}, w \Vdash \gamma_2$. We proceed as follows: (A′) iff $\mathfrak{M}, w \Vdash [\Uparrow \alpha]\varphi \boxdot\!\!\rightarrow [\Uparrow \alpha]\psi$ (i.e., vacuously true) iff $\mathfrak{M}, w \Vdash \gamma_2$ by (ii). Second, let us go to the case (∼ A′). For our goal, it suffices to show the equivalence between (B′) and $\mathfrak{M}, w \Vdash \gamma_1$. By (∼ A′) and the definition of $\$_w^{\Uparrow\alpha}$, (B′) is equivalent to:

$$(\exists S \in \$_w)(S \cap [\![[\Uparrow \alpha]\varphi]\!] \neq \emptyset \text{ and } S \cap [\![[\Uparrow \alpha]\varphi]\!] \subseteq [\![[\Uparrow \alpha]\psi]\!]).$$

By (ii), this is equivalent to $\mathfrak{M}, w \Vdash \gamma_1$, as required.

Therefore, we have shown ((A′) or (B′)) iff $\mathfrak{M}, w \Vdash \gamma_1 \vee \gamma_2$. □

3.4 A Complete Hilbert-style Axiomatization for Dynamified Hybrid Counterfactual Logic

Definition 3.3. Define $\mathbf{V}_{\mathcal{DHC}(@)}$ as the Hilbert-style axiomatization $\mathbf{V}_{\mathcal{DHC}(@)}$ extended with all the axioms in Table 2 and Table 3.

Then, $\mathbf{V}_{\mathcal{DHC}(@)}$ axiomatizes all the valid formulas in all systems of spheres in the following sense:

Theorem 2. *$\mathbf{V}_{\mathcal{DHC}(@)}$ is sound and complete with respect to all finite sphere models, i.e., for any $\varphi \in \mathsf{FORM}_{\mathcal{DHC}(@)}$, φ is valid on all finite sphere models iff φ is a theorem of $\mathbf{V}_{\mathcal{DHC}(@)}$. Therefore, $\mathbf{V}_{\mathcal{DHC}(@)}$ is decidable.*

[5] A key observation for the left-to-right direction is: if we find our witness P of (B′) from $\{\, ([\![\alpha]\!] \cap \bigcup \$_2) \cup S \mid S \in \$_w \,\}$, then this case and our assumption (i) implies that

$$\Big(\bigcup \$_w \cap [\![\alpha]\!]\Big) \cap [\![[\Uparrow \alpha]\varphi]\!] \neq \emptyset \text{ and } \Big(\bigcup \$_w \cap [\![\alpha]\!]\Big) \cap [\![[\Uparrow \alpha]\varphi]\!] \subseteq [\![[\Uparrow \alpha]\psi]\!].$$

This is because we have $(\bigcup \$_w \cap [\![\alpha]\!]) \cap [\![[\Uparrow \alpha]\varphi]\!] \subseteq ((\bigcup \$_w \cap [\![\alpha]\!]) \cup S) \cap [\![[\Uparrow \alpha]\varphi]\!] \subseteq [\![[\Uparrow \alpha]\psi]\!]$.

Proof. Fix any $\varphi \in \mathsf{FORM}_{\mathcal{DHC}(@)}$. We show the left-to-right direction alone, since it is easy to establish the soundness direction (the right-to-left direction) by Propositions 3 and 6. Suppose that φ is valid on all finite sphere models. By the reduction axioms in Table 2 and Table 3, we can find some formula $\varphi' \in \mathsf{FORM}_{\mathcal{HC}(@)}$ of the static language such that $\varphi' \leftrightarrow \varphi$ are valid on all sphere models. Thus, φ' is also valid on all finite sphere models. By Theorem 1, we can state that φ' is a theorem of $\mathbf{V}_{\mathcal{HC}(@)}$, hence a theorem of $\mathbf{V}_{\mathcal{DHC}(@)}$. Since all the axioms of Table 2 and Table 3 are axioms of $\mathbf{V}_{\mathcal{DHC}(@)}$, $\varphi \leftrightarrow \varphi'$ is a theorem of $\mathbf{V}_{\mathcal{DHC}(@)}$, which implies that φ is a theorem of $\mathbf{V}_{\mathcal{DHC}(@)}$, as desired. □

4 Further Directions

In terms of epistemic plausibility model (recall Remark 2.1), sphere models require the connectedness of $\leq_w$ (cf. Lewis (1981)). Can we drop this assumption from our study of dynamified hybrid counterfactual logic? As for the static logic, this question leads us to consider a hybridization of Veltman-Burgess-style minimal conditional logic (Burgess 1981, Veltman 1985). An interesting question should be: Is it possible to generalize our notion of (hybrid) sphere-bisimulation to the setting without the connectedness or the minimality assumptions?

We have shown that our salience update respects sphere-bisimulations in Proposition 4. As a next step, we hope to show that radical upgrade also respects sphere-bisimulation (possibly with some modification) [6]. It is also interesting to see if the notion of conservative upgrade $[\uparrow \alpha]$ (van Benthem 2007) fits our semantics based on sphere models. If so, does conservative upgrade respect sphere-bisimulations?

Acknowledgements I wish to thank Johan van Benthem and Davide Grossi for their interesting and clear lectures on dynamic epistemic logic at January Project 2010 when I visited ILLC. The participants of the ILLC seminar on Logics for Dynamics of Information and Preference 2010 have also provided

[6] In the final stage of writing up this paper, the author realized that our radical upgrade respects (in the sense of Proposition 4) the notion of sphere-bisimulation satisfying the following further requirements: (R-forth) If wZw' then $\bigcup \$_w \subseteq Z[\bigcup \$'_{w'}]$ and (R-back) If wZw' then $\bigcup \$'_{w'} \subseteq Z^{-1}[\bigcup \$_w]$. Recall from the description just before Remark 2.1 that we have defined the binary relation R on W by wRv iff $v \in \bigcup \$_w$. Then, the above requirements correspond exactly to the ordinary back and forth clauses for the notion of bisimulation for modal logic (cf. van Benthem (2010)). Therefore, we can regard (R-forth) and (R-back) as natural requirements to the notion of sphere bisimulation.

useful feedbacks to my presentation. I also would like to thank Tomoyuki Yamada. A discussion with him about the update using link-cutting on March 2011 led me to recall *permissibility kinematics* in Lewis (1979) and realize the notion of salience update in this paper. Finally, I wish to thank Yurie Hara for the discussion and her checking my English. However, any flaws are mine.

References

C. Areces and B. ten Cate. Hybrid logics. In P. Blackburn, J. van Benthem, and F. Wolter, editors, *Handbook of Modal Logic*, pages 821–868. Elsevier, 2007.

F. Baader and C. Lutz. Description logic. In P. Blackburn, J. van Benthem, and F. Wolter, editors, *Handbook of Modal Logic*, pages 757–820. Elsevier, 2007.

P. Blackburn. Arthur Prior and hybrid logic. *Synthese*, 150(3):329–372, 2006.

P. Blackburn and J. van Benthem. Modal logic: A semantic perspective. In P. Blackburn, J. van Benthem, and F. Wolter, editors, *Handbook of Modal Logic*, pages 1–84. Elsevier, 2007.

J. P. Burgess. Quick completeness proof for some logics of conditionals. *Notre Dame Journal of Formal Logic*, 22(1):76–84, 1981.

L. Demey. Some remarks on the model theory of epistemic plausibility models. *CoRR*, abs/1003.2790, 2010.

D. Lewis. *Counterfactuals*. Blackwell Publishing, 2 edition, 1973. 2nd edition was published in 1986.

D. Lewis. A problem about permission. In E. Saarinen, R. Hilpinen, I. Niiniluoto, and M. Hintikka, editors, *Essays in Honor of Jaakko Hintikka*, pages 163–175. Reidel, Dordrecht, 1979.

D. Lewis. Ordering semantics and premise semantics for counterfactuals. *Journal of Philosophical Logic*, 10(2):217–234, 1981.

O. Roy. A dynamic-epistemic hybrid logic for intentions and information changes in strategic games. *Synthese*, 171(2):291–320, 2009.

K. Sano. Hybrid counterfactual logics: David Lewis meets Arthur Prior again. *Journal of Logic, Language and Information*, 18(4):515–539, 2009.

B. ten Cate. *Model theory for extended modal languages*. PhD thesis, University of Amsterdam, Institute for Logic, Language and Computation, 2005.

J. Ulrik Hansen. A hybrid public announcement logic with distributed knowledge. To appear in Electronic Notes in Theoretical Computer Science 2011. Post-Proceedings of the International Workshop on Hybrid Logic and Applications (HyLo 2010).

J. van Benthem. Dynamic logic of belief revision. *Journal of Applied Non-Classical Logics*, 17(2):129–155, 2007.

J. van Benthem. *Modal Logic for Open Minds*. Center for the Study of Language and Information - Lecture Notes. Stanford Univ Center for the Study, 2010.

J. van Benthem and F. Liu. Dynamic logic of preference upgrade. *Journal of Applied Non-Classical Logics*, 17:157–182, 2007.

J. van Benthem and S. Minicã. Toward a dynamic logic of questions. In *LOGIC, RATIONALITY, AND INTERACTION*, volume 5834/2009 of *Lecture Notes in Computer Science*, pages 27–41, 2009.

H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*. Springer, 2008.

F. Veltman. *Logics for Conditionals*. Dissertation, Philosophical Institute, University of Amsterdam, 1985.

T. Yamada. Commands and changing obligations. In *Proceedings of the Seventh International Workshop on Computational Logic in Multi-Agent Systems (CLIMA VII)*, pages 1–19. Springer-Verlag, 2006.

Comparing Strengths of Beliefs Explicitly

Dick de Jongh and Sujata Ghosh

University of Amsterdam, University of Groningen
D.H.J.deJongh@uva.nl, sujata@ai.rug.nl

Abstract

Formulas $p \succ_B q$ and $p \succcurlyeq_B q$ are introduced in the existing logical framework for discussing beliefs to express that the strength of belief in p is greater than (or equal to) that in q. Besides its usefulness in studying the properties of the concept of greater strength of belief itself this explicit mention of the comparison in the logical language aids in defining several other concepts in a uniform way, viz. older and rather clear concepts like the operators for universality (the totality of possibilities considered by an agent), together with newer notions like plausibility (in the sense of 'more plausible than not') and disbelief. A major role is played in our investigations by the relationship between the standard plausibility ordering of the worlds and the strength of belief ordering. We keep the relation between the two orderings as light as possible to construct a system that allows for widely different interpretations. Finally, we provide an extension of the framework to the multi-agent setting, and we discuss the possibilities of extending our system to a dynamic one.

1 Introduction

Being subject to doubts and dilemmas while making decisions is like second nature to the human mind. The difference in the strengths of beliefs of an

agent regarding the occurrence of different events may clear doubts of this kind. In betting on games, people make their choices for putting their money on different teams, based on their strengths of beliefs about which team will win. Similarly, when voting, one's preference for the candidates is again based on the strength of beliefs about one candidate's ability to perform compared to the others. Thus, this notion is inherently present in various fields of research like decision theory, game theory and others.

Before proceeding further, let us first consider the following real life situation where comparison of strength of beliefs plays a key role in decision-making for recruitments.

Alice has applications for jobs in her departmental store. The first time Burt and Cora apply. Alice believes both can do the job, but her belief in Cora being able to do it is stronger than that Burt will be able to do it. She chooses Cora.

The second time Deirdre and Egon apply. She believes that Egon can do the job whereas she is ambiguous about Deirdre: she neither has the belief that Deirdre can do it, nor that she cannot. She chooses Egon.

The third time Fiona and Gregory apply. About both she is ambiguous, but her strength of belief in Gregory being able to do it is stronger than that in Fiona. She chooses Gregory, maybe she has to help him along a little.

The fourth time the applicants are Harold and Irma. She believes neither can do the job. She decides not to take one of those two and hold another round of applications. When she finds out that time is too short for that she thinks again, decides that she believes even less in Irma being able to do it than in Harold, and she takes Harold.

Let us point out one possible misunderstanding. Alice does not judge how well she thinks the applicants will perform, she just judges whether they will be able to do the job or not, a simple yes-no question. Of course, to combine our set up with beliefs in graded abilities would be highly interesting but that is a matter for future work.

All these situations regarding the belief states of Alice can be aptly described, if we talk not only about her beliefs but also compare the strength of her beliefs in the applicants. One sees here how a stronger belief can induce a preference.

One can argue that these situations can be described by the well-studied notion of *preference*, but the essence of describing the mental states of Alice will be lost then. This paper adds a new notion to this line of work, viz. comparing the strengths of beliefs, and very pertinently, doing this in a qualitative manner. The relationship with preference will be developed somewhat further in subsection 3.3.

The introduction of explicit notions of ordering for comparing strengths of

beliefs in the logical language has various applications. Besides its usefulness in studying the properties of the concept of greater strength of belief itself it aids in defining several other concepts in a uniform way. In models concerning knowledge (epistemic logic) equivalence classes of worlds (and in case of one agent: the set of all worlds) are naturally given by the indistinguishability relation connected with the knowledge operator. In models concerning belief (doxastic logic) a universality operator U is often introduced with the same purpose. In our set up this usually somewhat vague operator can be defined in terms of the order, the idea that the worlds the agent considers are the ones she considers possible in some manner is made explicit. In the semantics, the question - which worlds are going to be a part of the model, gets in our approach a clearer formal and intuitive understanding. It also becomes more evident that the universality operator must not be identified with the knowledge operator even if they both share the $S5$-properties.

Additionally, newer notions like plausibility (in the sense of 'more plausible than not') and disbelief can also be expressed. Above all it has its advantages in an explicit study of the properties of the orderings themselves, semantically and axiomatically. All these investigations can be carried over to a dynamic setting (Gerbrandy (1999), see also van Benthem (2007)), but we leave this for future work.

1.1 Related work

In Segerberg (1971), Gardenfors (1975), orderings of formulas are considered but their interpretations are probabilistic in nature. A binary sentential operator is introduced in the language with the intended interpretation 'at least as probable as'. While Gardenfors (1975) takes the explicit ordering operator in a simple language consisting of the truth-functional connectives only, Segerberg (1971) discusses this issue in a modal setting. As the interpretation suggests, the semantics is based on probability measures over worlds.

The notion of epistemic entrenchment Gärdenfors and Makinson (1988) gives a syntactic ordering of formulas, which is studied in connection with belief revision. The ordering influences the abandoning and retaining of formulas when a belief contraction or revision takes place.

Ordering of worlds provide an intuitive way to model various kinds of logical operators, specially the epistemic ones. To mention a few, Lewis Lewis (1973) proposed a plausibility ordering of worlds to provide a semantics for the counterfactual statements. With the goal of representing qualitative frameworks of belief in terms of the corresponding probabilistic ones, Spohn defines

a plausibility ordering of possible worlds in terms of ordinal functions Spohn (1988). More recently, such orderings have been discussed in the economics literature Board (2004).

Our focus lies on giving a qualitative framework for differing strengths of beliefs that an agent may have on different propositions (possibly, individuals), influencing her decision making process. Semantically, rather than modeling in terms of *world ordering*, we rely on *set ordering* for comparing belief strengths.

We should mention here that the idea of modeling epistemic notions in terms of set orders is not really new. In Halpern (1997), preferential structures are considered to give semantics to a logic of relative likelihood. A preference ordering over worlds is lifted to an ordering of sets of worlds. Plausibility measures over sets of worlds are considered in Friedman and Halpern (2001) to give a semantics of default logic. These measures induce a set ordering which provides an interpretation of the notion of belief (similar to our notion of plausibility in Section 3.1) . For a detailed overview, see Halpern (2003).

While our interpretation of belief is given in terms of world ordering, a set ordering is used to interpret strength of belief. This distinguishes our work form the ones mentioned above. Moreover, this ordering of sets of worlds is only partly determined by the ordering of the worlds. We discuss our reasons for this in later sections, especially in Section 6.

2 Comparing strengths of beliefs explicitly

Possible-world semantics Kripke (1963) has been used to model knowledge as well as belief. An extensive discussion together with all pre-requisite definitions can be found in Halpern and Moses (1992). In this work we are only concerned with *beliefs* of agents, comparison of their strengths as well as some related notions like *universality*, *safe beliefs*, *plausibility*, *disbelief* and others. Various debates and discussions are still going strong among the philosophers regarding the axioms that characterize belief - for this paper we will stick to the *KD45*-model of belief.

In the following, we talk about Kripke structures as well as the plausibility models van Benthem (2007), Baltag and Smets (2008) as and when needed while talking about beliefs. The readers should note that plausibility models are more general in nature in the sense that one can always build up a *KD45* Kripke structure from them, as described in Baltag and Smets (2008).

With this brief overview, we now move on to introduce explicit ordering of beliefs in the logical language, which is the essential new feature of this paper.

This explicit mention of such comparison of beliefs provides an informative and uniform way to discuss relevant issues like disbeliefs, plausibility and others.

To introduce this comparison of strengths of beliefs explicitly in the logical language, we add new relation symbols to the existing modal language of belief to form the language of *Belief logic with explicit ordering* (*KD45−O*), whose language is defined as follows:

Definition 2.1. Given a countable set of atomic propositions Φ, formulas φ are defined inductively:

$$\varphi := \bot \mid p \mid \neg\varphi \mid \varphi \vee \varphi \mid B\varphi \mid \varphi \succcurlyeq_B \psi \mid$$

where $p \in \Phi$.

The intuitive reading of the formula $B\varphi$ is "φ is believed", and that of $\varphi \succcurlyeq_B \psi$ is "belief in φ is at least as strong as belief in ψ". We introduce the notations $\varphi \succ_B \psi$ for $(\varphi \succcurlyeq_B \psi) \wedge \neg(\psi \succcurlyeq_B \varphi)$ and $\varphi \equiv_B \psi$ for $(\varphi \succcurlyeq_B \psi) \wedge (\psi \succcurlyeq_B \varphi)$. Intuitively, they can be read as "belief in φ is stronger than that in ψ" and "belief in φ and ψ are of same strength", respectively. We now define a model for this logic.

Definition 2.2. A *KD45−O* model is defined to be a structure $\mathcal{M} = (S, \leq, \geq_B, V)$, where S is a non-empty finite set of states, V is a valuation assigning truth values to atomic propositions in states, $\leq$ is a quasi-linear[1] order relation (a plausibility ordering) over S, and $\geq_B$ is a quasi-linear order relation over $\mathcal{P}(S)$, satisfying the conditions

(1) If $X \subseteq Y$, then $Y \geq_B X$

(2) If $\mathcal{B}$ is the set of all $\leq$-minimal worlds (the set of most-plausible worlds, called the *center*), then $\mathcal{B} \subseteq X$ and $\mathcal{B} \nsubseteq Y$ imply $X >_B Y$, where $X >_B Y$ iff $X \geq_B Y$ and not $(Y \geq_B X)$.

(3) If X is non-empty, then $X >_B \emptyset$.

The first condition says that larger sets of worlds are at least as plausible, the second one that the sets containing the center are more plausible than those not containing it, and the third one that non-empty sets are more plausible than the empty set. Truth on the center suffices to make an assertion to be believed. Note that all the models are considered to be finite. This assumption ensures

[1] A binary relation $\leq$ on a non-empty set S is said to be quasi-linear if it is reflexive, transitive and linear, i.e. a total pre-order. That we do take the order to be quasi-linear, but not more generally a pre-order is not a matter of principle but rather of convenience.

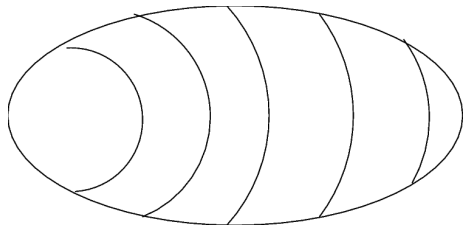


Figure 1: Plausibility ordering

the existence of minimal worlds in terms of the plausibility ordering of the model. The truth definition for formulas φ in a *KD45−O* model $\mathcal{M}$ is as usual with the following clauses for the belief and ordering modalities.

$\mathcal{M}, s \models B\varphi$ iff $\mathcal{M}, t \models \varphi$ for all $\leq$-minimal worlds t.

$\mathcal{M}, s \models \varphi \geqslant_B \psi$ iff $\{t \mid \mathcal{M}, t \models \varphi\} \geq_B \{t \mid \mathcal{M}, t \models \psi\}$.

We consider $\geqslant_B$ to be a global notion, if $\varphi \geqslant_B \psi$ is true anywhere in the model, it is true everywhere. So, it is either true or false throughout the whole model; $\geqslant_B$ is a global notion like B. Of course, being global in the model is strongly connected with introspection. In general we support the idea that the agent knows everything about the model except which state represents the actual world. From the definition of $>_B$, it follows that,

$\mathcal{M}, s \models \varphi >_B \psi$ iff $\{t \mid \mathcal{M}, t \models \varphi\} >_B \{t \mid \mathcal{M}, t \models \psi\}$.

Thus, $>_B$ is also a global notion. We will now show that the universal modality U can also be expressed in the language. The modality $E\varphi$ (the abbreviated form of $\neg U \neg \varphi$) can be defined as $\varphi >_B \bot$, and hence $U\varphi$ ($= \neg E \neg \varphi$) itself as $\bot \geqslant_B \neg\varphi$: $U\varphi$ expresses that φ is true in all possible worlds in the model, whereas $E\varphi$ stands for existence of a possible world in the model where φ is true. The formula $\varphi >_B \bot$, which defines $E\varphi$, expresses the intuition that those worlds should be considered in the model of which the existence is expressed by a positive strength of belief, those possibilities which the agent does not want to exclude. Evidently, we have,

$\mathcal{M}, s \models U\varphi$ iff $\mathcal{M}, t \models \varphi$ for all worlds t.

Alice's belief states (as described in the introduction) can now be formally presented as follows: suppose each of the applicants' names denotes the proposition that "he (she) can do the job". Cora $\succ_B$ Burt in the first case; B(Egon) $\wedge$ ($\neg$B(Deirdre) $\wedge$ $\neg$B($\neg$Deirdre)) implies that Egon $\succ_B$ Deirdre in the second case, with the third case simply being Gregory $\succ_B$ Fiona again, and the fourth one, B($\neg$Harold)$\wedge$B($\neg$Irma), and later Harols $\succ_B$ Irma. The readers can easily see that in the second case there is some reasoning going on which leads to Egon being given the job, because Alice's belief in the ability of Egon is stronger than her belief in the ability of Deirdre.

2.1 Axioms and Completeness

In this subsection we introduce the proof system *KD45−O*, and discuss its relationship to the *KD45−O*-models. The system consists of the following axioms and rules.

Definition 2.3. The system $KD45 - O$ consists of

a) all *KD45* axioms and rules for B

b) ordering axioms:

$\varphi \succcurlyeq_B \varphi$ (refl-axiom)

$(\varphi \succcurlyeq_B \psi) \wedge (\psi \succcurlyeq_B \chi) \rightarrow \varphi \succcurlyeq_B \chi$ (trans-axiom)

$(\varphi \succcurlyeq_B \psi) \vee (\psi \succ_B \varphi)$ (lin-axiom)

$(B\varphi \wedge \neg B\psi) \rightarrow (\varphi \succ_B \psi)$ (center-axiom)

$(\varphi \succcurlyeq_B \psi) \rightarrow B(\varphi \succcurlyeq_B \psi)$ (intros-axiom1)

$(\varphi \succ_B \psi) \rightarrow B(\varphi \succ_B \psi)$ (intros-axiom2)

$\bot \succcurlyeq_B \neg(\varphi \rightarrow \psi) \rightarrow (\psi \succcurlyeq_B \varphi)$ ($U \succcurlyeq_B$-axiom)

$\varphi \rightarrow (\varphi \succ_B \bot)$ (existence axiom)

$(B\varphi \succ_B \bot) \rightarrow B\varphi$ (un.center-axiom)

c) inclusion rule:

$$\frac{\varphi \rightarrow \psi}{\psi \succcurlyeq_B \varphi}$$ (inclusion rule)

Let us discuss these axioms, and, more or less informally, their soundness for the models introduced above. On the way, we will make clear that the $S5$-properties of the universal modality U Goranko and Passy (1992) are all provable in $KD45 - O$. We will discuss possible additional axioms in the subsection immediately after this one.

Since, by the set-ordering relation in the $KD45{-}O$ model, $\geq_B$ is a reflexive, transitive and connected relation over $\mathcal{P}(S)$, and $>_B$ is the corresponding strict ordering, the three basic ordering axioms are evidently sound. From these axioms it immediately follows that $>_B$ is a transitive relation and also that $U\psi$ and $\neg E\neg\psi$ are provably equivalent.

The soundness of the inclusion rule follows from the condition (1) that larger sets in the models are at least as plausible as smaller sets. It has two important consequences. The first one is the *equivalence rule*:

$$\frac{\varphi \leftrightarrow \psi}{\varphi \equiv_B \psi}$$

which implies that substituting logically equivalent formulas for each other in ordering formulas leads to logically equivalent formulas. Since the rest is modal logic, this is the only thing needed to show that substituting them for each other anywhere leads to logically equivalent formulas. The second important consequence is the necessitation rule for U (Ugen-rule). Because, if $\vdash_{KD45-O} \varphi$, then $\vdash_{KD45-O} \neg\varphi \rightarrow \bot$. The inclusion rule now gives $\vdash_{KD45-O} \bot \geqslant_B \neg\varphi$, i.e., $\vdash_{KD45-O} U\varphi$.

Three axioms involve U and/or E. We have the $U \geqslant_B$-axiom, which can be reformulated as

$$U(\varphi \rightarrow \psi) \rightarrow (\psi \geqslant_B \varphi),$$

and which is also connected to condition (1). In fact, it expresses that formulas that are equivalent in the model when replacing each other lead to formulas that are equivalent in the model. Moreover, this axiom can be used to prove the K-axiom for U. For, assume we have $U\varphi$ and $U(\varphi \rightarrow \psi)$, but not $U\psi$, i.e., $\neg\psi >_B \bot$. Then, by the first assumption we have $\bot \geqslant_B \neg\varphi$, and hence $\neg\psi >_B \neg\varphi$.

By the second assumption, we have also $U(\neg\psi \rightarrow \neg\varphi)$ (equivalence!), and using the $U \geqslant_B$-axiom, $\neg\varphi \geqslant_B \neg\psi$, a contradiction.

Next we have the existence axiom, which can be reformulated as

$$\varphi \rightarrow E\varphi$$

The existence axiom is basically the same as the ordered formula for $U\varphi \to \varphi$, one of the $S5$-axioms for U. The last of the three is the un.center axiom, which can be reformulated as

$$EB\varphi \to B\varphi \quad \text{(un.center-axiom)}$$

It derives from the fact that the $KD45{-}O$ models have a unique center $\mathcal{B}$. It makes B a global property: the principle $B\varphi \to UB\varphi$ readily follows by first proving $E\neg B\varphi \to \neg B\varphi$.

Since the ordering formulas are either globally true or globally false in the models, we have the soundness of the two introspection axioms:

$$(\varphi \succcurlyeq_B \psi) \to B(\varphi \succcurlyeq_B \psi)$$
$$(\varphi \succ_B \psi) \to B(\varphi \succ_B \psi)$$

It immediately follows that

$$\neg(\varphi \succcurlyeq_B \psi) \to B\neg(\varphi \succcurlyeq_B \psi)$$
$$\neg(\varphi \succ_B \psi) \to B\neg(\varphi \succ_B \psi)$$

The converses of all these implications above follow from the lin-axiom. This means that all these ordering statements can be considered to be B-statements, i.e. $\varphi \succcurlyeq_B \psi$, $\varphi \succ_B \psi$, $U\varphi$, $E\varphi$ are all B-statements (and remember that equivalent formulas can be replaced by each other modulo provable equivalence). As a result, the inclusion formula concerning the belief and the universal modality, viz. $U\varphi \to B\varphi$ also follows. And the $U\psi \to UU\psi$ and $\neg U\psi \to U\neg U\psi$ axioms for U follow as well; because of the very significant property of $U\psi$ being a B-statement the un.center-axiom applies to U-statements as well. We have now covered all the $S5$-axioms for U.

We are now ready to prove the following completeness theorem which is the most basic and important result of this work.

Theorem 1. *$KD45{-}O$ is sound and complete with respect to $KD45{-}O$ models.*

Proof. Soundness has been treated above. Moreover we will freely use U meaning its translation into $KD45-O$, and we can assume that U has the $S5$-properties. We will show completeness using finite sets of sentences.

Assume $\nvdash_{KD45-O} \varphi$. We will have to construct a counter-model to φ as a $KD45{-}O$-model. We take a finite *adequate* set Φ containing φ. In this case

an adequate set will be: a set of formulas that is closed under subformulas containing with each formula ψ (a formula equivalent to) $\neg\psi$, containing with $B\psi$ and $B\chi$ (a formula equivalent to) $B(\psi \wedge \chi)$ and a formula (equivalent to) $B(\psi \vee \chi)$. We also need Φ to contain with each formula $B\varphi$ the formula $UB\varphi$. Finally, Φ contains $B\top$ and $B\bot$. It is easy to see that any finite set is contained in a finite adequate set. We use the Henkin method restricted to Φ. Consider the m.c. (maximally consistent) subsets of Φ. In particular consider such an m.c. set Φ_0 containing $\neg\varphi$.

The relations $\mathcal{R}_B$ and $\mathcal{R}_U$ are defined as follows:

$$\begin{array}{lll} P\mathcal{R}_B Q & \text{iff} & \text{(1) for all } B\varphi \text{ in } P,\ \varphi \text{ as well as } B\varphi \text{ are in } Q, \\ & & \text{(2) for all } \neg B\varphi \text{ in } P,\ \neg B\varphi \text{ in } Q. \\ P\mathcal{R}_U Q & \text{iff} & \text{(1) for all } U\varphi \text{ in } P,\ \varphi \text{ as well as } U\varphi \text{ are in } Q, \\ & & \text{(2) for all } \neg U\varphi \text{ in } P,\ \neg U\varphi \text{ in } Q \end{array}$$

We have to show that $\mathcal{R}_U$ is an equivalence relation and $\mathcal{R}_B$ a Euclidean subrelation of $\mathcal{R}_U$. Finally, within one U-equivalence class there is one, nonempty set of B-reflexive elements, which forms a B-equivalence class. Since all these things are standard we skip this part.

We now take the submodel generated by $\mathcal{R}_U$ from Φ_0. The set of worlds W of our model will be the set of worlds in this submodel and the $\mathcal{R}_B$ and $\mathcal{R}_U$ the restrictions of the original $\mathcal{R}_B$ and $\mathcal{R}_U$ to this submodel. $\mathcal{R}_U$ is now the universal relation.

As before, we write $\mathcal{B}$ for the set of $\mathcal{R}_B$-reflexive elements. The axiom $B\varphi \rightarrow UB\varphi$ implies that this set is unique and a B-equivalence class. The world plausibility ordering is given as follows: any world in $\mathcal{B}$ is more plausible than any in $W \setminus \mathcal{B}$, and within these two sets, the worlds are equi-plausible. So, with respect to the modal operators B and U the model behaves properly, and we have a proper world-ordering as well. We will now have to order $\mathcal{P}(W)$ in a proper way.

Let us say that ψ *represents* subset X of W if X is the set of nodes where ψ is true, which we may write as $V(\psi) = X$. We say that X is *representable* if for some $B\psi$ in Φ, ψ represents X. By the conditions on Φ the representable sets are closed under unions and intersections, and contain W itself and the empty set.

The representable subsets of Φ are quasi-linearly ordered by the relation $\geq_1$ defined by $V(\psi) \geq_1 V(\chi)$ iff $\psi \succcurlyeq_B \chi$ is true in the model, $V(\psi) >_1 V(\chi)$ iff $\psi \succ_B \chi$ is true in the model. These follow from the first three ordering axioms.

Moreover, if $V(\psi) \subseteq V(\chi)$ then $V(\psi) \geq_1 V(\chi)$ (subset condition), by the axiom: $U(\chi \rightarrow \psi) \rightarrow \ \psi \succcurlyeq_B \chi$. Finally if $V(\psi)$ properly contains $\mathcal{B}$ and $V(\chi)$ does

not, then $V(\psi) >_1 V(\chi)$ (sufficient belief condition) by the axiom: $B\psi \wedge \neg B\chi \rightarrow \psi \succ_B \chi$.

So, $\geq_1$ behaves properly on the representable elements of $\mathcal{P}(W)$. What remains is to extend $\geq_1$ to an ordering $\geq$ with the right properties over all of $\mathcal{P}(W)$.

Take an arbitrary subset X of W. We define $R(X)$ to be the largest subset of X that is representable. That such a set exists follows from the fact that the representable subsets are closed under finite unions and the finiteness of the model.

We now define $X \geq Y$ iff $R(X) \geq_1 R(Y)$. This immediately makes $\geq$ a quasi-linear order. That $\geq$ satisfies the *subset condition* follows from the fact that, if $X \subseteq Y$, then $R(X) \subseteq R(Y)$.

We will conclude this proof with a lemma showing that $\mathcal{B}$ is representable, i.e. $\mathcal{B} = R(\mathcal{B})$. From that result it follows that, if $\mathcal{B} \subseteq X$, then $\mathcal{B} \subseteq R(X)$. This is clearly sufficient to ensure the *sufficient belief condition*. So, once we finish the proof of the following lemma, we are done.

Lemma 1. $\mathcal{B}$ is representable: *Consider w not in $\mathcal{B}$. Then it is not the case that $w\mathcal{R}_B w$. This means that, for some particular $B(\psi_w)$ in Φ, $B(\psi_w)$ is in w but ψ_w is not. (Other possibilities are excluded because we already know that $B(\psi_w)$ and $\neg B(\psi_w)$ true everywhere or nowhere.) Note that this implies that ψ_w is true all over $\mathcal{B}$. Consider the conjunction ψ of all ψ_w for w in the complement of $\mathcal{B}$. $B(\psi)$ is a member of Φ while ψ is true in all elements of $\mathcal{B}$, but is falsified at all elements u in the complement of $\mathcal{B}$, since ψ implies ψ_u and ψ_u is falsified in u. We have shown that $\mathcal{B}$ is represented by ψ.*

This completes the proof. □

Since the counter-model constructed is finite, we also have that the logic *KD45−O* is decidable.

2.2 Additional principles

Before ending this section we mention some formulas which we did not need as axioms, but are definitely worth thinking about as possible additions to *KD45 − O*. One of them is,

$$(\varphi \succ_B \bot) \rightarrow (\top \succ_B \neg\varphi),$$

which says that, if φ is true somewhere, then $\neg\varphi$ is less believable than a tautology. The other direction of the implication can be derived. An equivalent formulation is,

$$(\varphi \geqslant_B \top) \rightarrow (\bot \geqslant_B \neg\varphi).$$

To make this true, the model needs an extra clause, saying that,

$$\text{if } S \neq X \text{ then } S >_B X.$$

This seems a very reasonable addition as it makes the models more symmetric. Also $U\varphi$ can by its use be more simply defined as $\varphi \geqslant_B \top$.

An axiom that expresses another form of symmetry is

$$(B\neg\psi \wedge \neg B\neg\varphi) \rightarrow (\varphi >_B \psi).$$

In the presented system one can only get $\neg\psi >_B \neg\varphi$ from $B\neg\psi \wedge \neg B\neg\varphi$. In the final discussion we will use this axiom to define the world order in terms of the set order. A more general version that implies both previous possible additions is

$$(\varphi >_B \psi) \rightarrow (\neg\psi >_B \neg\varphi).$$

which, if considered, definitely increases the already-existing probabilistic flavor of the axiomatization. Another possible principle with a similar flavor is

$$(\varphi >_B \psi) \rightarrow (\varphi \wedge \neg\psi) >_B (\psi \wedge \neg\varphi).$$

This principle exemplifies the feeling that if φ is more believed than ψ, then that can only be based on the non-common parts of the extensions of φ and ψ: the common part of φ and ψ should be irrelevant in the estimation of their relative strengths of belief. Readers can note here that if we strengthen this formula to its bi-implication, then $(\varphi >_B \psi) \rightarrow (\neg\psi >_B \neg\varphi)$ follows.

3 Applying the explicit ordering framework

We now show that the explicit notions of ordering for comparing strengths of beliefs in the logical language aid in expressing several other related concepts in a uniform way, viz. plausibility, disbelief, and preference.

3.1 Plausibility

Comparing the strength of beliefs explicitly has its various advantageous applications. By plausibility of a proposition we generally mean that we tend to believe in its happening rather than its not happening. That is the interpretation we take here. Hence, in terms of ordered formulas, $P\varphi$ can be expressed as $\varphi \succ_B \neg\varphi$. Of course, there are other possible notions of plausibility, but here we interpret $P\varphi$ as 'more plausible than not'. We now explore this notion of 'plausibility' in terms of belief ordering.

An important principle that will be valid for the *plausibility* operator P is $U(\varphi \to \psi) \to (P\varphi \to P\psi)$. This holds because if $U(\varphi \to \psi)$, not only $\psi \succcurlyeq_B \varphi$, but $U(\varphi \to \psi)$ implies $U(\neg\varphi \to \neg\psi)$, so also $\neg\varphi \succcurlyeq_B \neg\psi$. So, if $P\varphi$, i.e., $\varphi \succ_B \neg\varphi$, then $\psi \succcurlyeq_B \neg\varphi$, so $\psi \succ_B \neg\psi$, i.e., $P\psi$. This principle leads to consequences like $P(\varphi \wedge \psi) \to P\varphi$.

The reason to take the set semantics for ordering formulas (cf. Definition 2.2) becomes clear. If we would adhere to the semantics we may have had for $\succ_B$ in terms of plausibility ordering for worlds (instead of sets of worlds), $P\varphi$ would become equivalent to $B\varphi$, which obviously is undesirable.

One can just subdivide the most plausible worlds (the center) into more and less plausible ones to rectify this, but besides endangering the transition to dynamics this will not yet be really satisfactory in its own right. It will result in interpreting $P\varphi$ into something like 'φ is weakly believed'. This would make the modal logic of P a normal modal logic (of weak belief). In particular $P\varphi \wedge P\psi \to P(\varphi \wedge \psi)$, which is equivcalent to the K-axiom, would become valid, which is not very intuitive.

For example, you may judge it more plausible than not that your next client will be male. Similarly, you may consider it to be plausible that your next client will be a foreigner. But, it doesn't follow that it is more plausible than not that the next client will be a foreign male, most of one's foreign clients may be female.

We now move on to showing an independent axiomatization of the plausibility logic P. The language of the P-logic is given by

$$\varphi := p \mid \neg\varphi \mid \varphi \vee \varphi \mid P\varphi$$

We read $P\varphi$ as "φ is plausible". As mentioned above, the intuitive meaning of $P\varphi$ can be captured by the formula $\varphi \succ_B \neg\varphi$, and as such, the truth definition of $P\varphi$ in the *KD45–O* model is given by,

$$\mathcal{M}, s \models P\varphi \text{ iff } \{t \mid \mathcal{M}, t \models \varphi\} >_B \{t \mid \mathcal{M}, t \models \neg\varphi\}.$$

Theorem 2. *P-logic is complete and its validities are completely axiomatized by the following axioms and rules:*

(a) all propositional tautologies and inference rules

(b) plausibility axioms:

$P\psi \wedge P\varphi \rightarrow P(\psi \wedge P\varphi)$

$\neg P\varphi \rightarrow P\neg P\varphi$

$P\varphi \rightarrow \neg P\neg\varphi$

$P\top$

c) monotonicity rule:

if $\varphi \rightarrow \psi$ *then* $P\varphi \rightarrow P\psi$

Before giving the proof let us make some remarks on the axioms. The monotonicity rule implies the necessitation rule $\varphi / P\varphi$, by the axiom $P\top$. Moreover, from the monotonicity rule the equivalence rule,

if $\varphi \leftrightarrow \psi$ then $P\varphi \leftrightarrow P\psi$

immediately follows. This, in its turn means that provable equivalents can be substituted for each other without impairing provability. As a second plausibility axiom one might have expected $P\psi \wedge \neg P\varphi \rightarrow P(\psi \wedge \neg P\varphi)$, but this follows from our second axiom. To see this, just note that $\neg P\psi$ and $P\neg P\psi$, by use of our second and third axiom, are provably equivalent.

Proof. First of all, we show that any formula in P-logic is equivalent to a formula with P-depth at most one. For that purpose we first derive the following schemes:

(1) $P\psi \rightarrow (\varphi \leftrightarrow \varphi[\top / P\psi])$

(2) $\neg P\psi \rightarrow (\varphi \leftrightarrow \varphi[\bot / P\psi])$

Here, $\varphi[\top / P\psi]$ means φ with $\top$ substituted for some occurrences of $P\psi$. We prove by induction on the complexity of formulas φ with possible occurrences of $\top$ and $\bot$.

In the base case, that is for the atomic propositions, propositional constants and $P\psi$, the result follows immediately.

Induction step. This is trivial for the boolean connectives. So, it suffices to prove it for $P\varphi$ assuming it holds for φ. From the induction hypothesis for the first scheme it follows that $(P\psi \wedge \varphi) \leftrightarrow (P\psi \wedge \varphi[\top/P\psi])$ is provable. Now assume $P\psi$ and $P\varphi$. By an axiom $P(\varphi \wedge P\psi)$ follows. From the fact just proved it follows that $P(\varphi[\top/P\psi] \wedge P\psi)$ and hence $P(\varphi[\top/P\psi])$. The proof for the second scheme is very similar.

To see that these schemes imply that each formula in P-logic is equivalent to a formula with P-depth at most one, just note that $\vdash \varphi \leftrightarrow ((P\psi \wedge \varphi) \vee (\neg P\psi \wedge \varphi))$. Now, if we want to get rid of occurrences of $P\psi$ in φ we can replace φ by $((P\psi \wedge \varphi[\top/P\psi]) \vee (\neg P\psi \wedge \varphi[\bot/P\psi]))$. By doing this consecutively for all occurrences of some $P\psi$ with no occurrences of P in ψ we obtain the desired result.

Next, we show that any consistent set has a model. Assume we have a consistent set in the P-logic which can be extended to a maximal consistent set Γ, say. Since we can restrict attention to formulas which are boolean combinations of atoms and formulas of the form $P\varphi$ where φ no longer contains P, a maximal consistent set is essentially only a set of atoms, negations of atoms, and such $P\varphi$'s and $\neg P\varphi$'s.

Let us just take a finite number of atoms to keep things finite, and let us take a maximal consistent set Γ of the form described above. We now make a model in our sense where $P\varphi$ gets interpreted as $\varphi >_B \neg\varphi$. The worlds will be simply defined by a number of atoms being true in it and the rest of the atoms false. Let us now consider the following model, $\mathcal{M} = (S, \leq, \geq_B, V)$, where S is the set of all such worlds. The ordering of the subsets is as follows: There are 5 equivalence classes in the ordering starting with the highest grade of believability. We take membership of those classes to determine the degree of belief in the sets. As representing formulas we just take purely propositional ones.
(1) The whole set, which is of course represented by $\top$ (or other tautologies).
(2) The sets represented by those φ for which $P\varphi$ is in Γ (except for $\top$).
(3) The sets represented by those φ for which $\neg P\varphi$ is in Γ as well as $\neg P\neg\varphi$.
(4) The sets represented by those φ for which $P\neg\varphi$ is in Γ (except for $\bot$).
(5) The empty set, which is of course represented by $\bot$. These are all possibilities because of axiom $P\varphi \rightarrow \neg P\neg\varphi$. Finally we take $\mathcal{B}$, the center, to be the whole set (so, there are no beliefs except the trivial one in $\top$).

The two things we have to check are: First, that, if a set is in class (2), then any larger one will be in (2) as well (or in (1)). This follows from the *monotonicity rule*, since by the fact that the worlds are determined by the atoms true in them all inclusions are logical inclusions. Similarly for the other classes. Second, that, if a set X contains all of $\mathcal{B}$, and another set Y doesn't, then $X > Y$. That is trivial: X has to be $\mathcal{B}$, the whole set, and Y isn't.

Finally, we see that Γ is satisfied by the world in the model that makes exactly its atoms true. So, for each consistent set we can have a model in *KD45−O*. So, the axioms and rules given in Theorem 2.6 axiomatize the *P*-logic of 'more plausible than not'. It is also worth-mentioning why $(P\varphi \wedge P\psi) \rightarrow P(\varphi \wedge \psi)$ will fail in general. There may be sets in (2), the intersection of which, is not in (2). □

Evidently, $P\varphi$ is a global notion - its value does not vary through the model. Again, *P* is clearly an introspective notion.

Let us finally note that an interpretation of $P\varphi$ as φ having probability more than 0.5 (or any other number between 0.5 and 1) leads to exactly the *P*-axioms, provided one considers the probability statements themselves to always have probability 1.

Neighborhood models

It is good to mention that a different, more standard, but equivalent semantics for the *P*-logic exists: neighborhood models Chellas (1980). A *neighborhood frame* consists of a set of worlds W and a function ν that maps each world w onto a set of subsets of W such that, if $X \in \nu(w)$ and $X \subseteq Y$, then $Y \in \nu(w)$. A *neighborhood model* is a neighborhood frame with a valuation as usual. A formula $P\varphi$ will be true in w if $V(\varphi) \in \nu(w)$.

It is clear that in our case the set of worlds S together with the (constant) function ν that maps each world to $\{X | X >_B S - X\}$ is a neighborhood frame. The corresponding neighborhood models will give exactly the same truth conditions as our models. The special properties that the neighborhood frames for the *P*-logic have beyond the standard ones mentioned above are:

- The function ν is constant on W,
- If $X \in W$, then $S - X \notin W$,
- $\nu(w)$ is non-empty, it contains W.

The logics corresponding to the neighborhood frames are called *monotonic* logics Hansen (PP-2003-24). The minimal monotonic logic has beyond propositional logic just the axiom $P\top$, and the monotonicity rule.

Belief and plausibility

We now consider a system having both belief and the plausibility operator, viz. the *BP*-system. This system will provide pointers to discuss logics of belief and disbelief in the next subsection. The language is that of the *P*-logic, together with the additional modal operator for belief, *B*.

$$\varphi := p \mid \neg\varphi \mid \varphi \vee \varphi \mid P\varphi \mid B\varphi$$

Some validities of this logic in the *KD45−O* model are,

- $B\varphi \to P\varphi$
- $P\varphi \to BP\varphi$
- $\neg P\varphi \to B\neg P\varphi$

Theorem 3. *BP-logic is complete and its validities are completely axiomatized by the following axioms and rules:*

a) all propositional tautologies and inference rules

b) all KD45 axioms and rules

c) all P axioms and rules

d) special axioms:

$B\varphi \to P\varphi$
$P\varphi \to BP\varphi$

The proof is very similar to that for the *P*-logic. It starts with proving that the axioms force all formulas to be equivalent to boolean combinations of atoms and formulas of the form $P\varphi$ and $B\varphi$, where φ is boolean. Analogously to the *P*-logic one first has to prove

(1) $P\psi \to (\varphi \leftrightarrow \varphi[\top/P\psi])$

(2) $\neg P\psi \to (\varphi \leftrightarrow \varphi[\bot/P\psi])$

(3) $B\psi \to (\varphi \leftrightarrow \varphi[\top/B\psi])$

(4) $\neg B\psi \to (\varphi \leftrightarrow \varphi[\bot/B\psi])$

One needs the theorem $B\varphi \rightarrow PB\varphi$, which follows immediately from $B\varphi \rightarrow BB\varphi$ and $B\varphi \rightarrow P\varphi$. Instead of five grades of believability we will now get seven, e.g. the second class splits into sets represented by φ with $B\varphi$ true and the ones representable by φ with $\neg B\varphi$ and $P\varphi$ true.

It is noteworthy that the principle $B\varphi \wedge P\psi \rightarrow P(\varphi \wedge \psi)$ of Burgess (1969) fails in the BP-logic. It is not difficult to construct a counterexample.

3.2 Disbelief

Disbelief in a proposition is governed by exactly the opposite situation to the one discussed in the previous subsection, $D\varphi$ can be expressed as $\neg\varphi \succ_B \varphi$, that is $P\neg\varphi$.

With the huge amount of work going on in logics of belief and belief revision, consideration of disbelief as a separate epistemic category came to fore in the latter part of last decade (Ghose and Goebel (1998), Gomolinska (1998)). Consideration of changing or revising disbeliefs as a process analogous to belief revision was taken up by Gomolinska and Pearce (2001). Belief-disbelief pairs i.e. simultaneous consideration of belief and disbelief sets were also taken up (Chopra et al. (2002), Chakraborty and Ghosh (to appear)) through which various connections of possible inter-connectivity of beliefs and disbeliefs have come into focus. As mentioned earlier our notion of explicit belief ordering provides another path into expressing the concept of disbelief.

The basic idea for disbelieving a proposition is that the inclination to believe in its negation is stronger than that to believe it. Consequently, disbelieving is a much weaker notion than believing the negation of the proposition, but it should imply that one does not believe in the proposition. In other words, $D\varphi$ is implied by $B\neg\varphi$ and implies $\neg B\varphi$ but not the other way around in either case.

In general, if a person faces a decision based on whether a certain state of affairs is the case or an event happens, she may not have enough evidence to believe that the state of affairs is the case or is not the case. Then she may base her decision on whether she thinks the state of affairs plausible or disbelieves in it. Only in the case that her strength of belief in the two possibilities is equal, translated into our framework as $\varphi \equiv_B \neg\varphi$, it is a real tossup for her.

Various principles for the 'disbelief' operator together with the 'belief' one have been discussed in Gomolinska (1998) in the autoepistemic logic framework of Moore (1985). As such, the possible world semantics provided there which is based on separate sets of worlds for beliefs and disbeliefs is not very interesting, and suffers from 'disjointedness' as well as 'mirror-image' problems. These questions will not arise in the semantics we propose here. The basic reason is

the fact that 'disbelief' is given a global stance in contrast to 'belief' which is apparent from their respective interpretations. This also emphasizes the fact that disbelieving something is different from both 'not believing' as well as 'believing the negation'.

We now focus on getting a more feasible logic of belief and disbelief in similar lines to *BP*-logic introduced earlier. From our formal understanding $D\varphi$ is same as $P\neg\varphi$ and hence we get the following dual axiomatization of the *BD*-logic .

Theorem 4. *BD-logic is complete and its validities are completely axiomatized by the following axioms and rules:*

a) all propositional tautologies and inference rules

b) all KD45 axioms and rules

c) disbelief axioms:

$D\psi \wedge D\varphi \rightarrow D(\psi \vee \neg D\varphi)$

$\neg D\varphi \rightarrow DD\varphi$

$D\varphi \rightarrow \neg D\neg\varphi$

$D\bot$

d) special axioms:

$B\varphi \rightarrow D\neg\varphi$

$D\varphi \rightarrow BD\varphi$

e) anti-monotone rule:

if $\varphi \rightarrow \psi$ *then* $D\psi \rightarrow D\varphi$.

The proof follows similarly as in the case of *BP*-logic. Some interesting validities of this logic are,

- $B\neg\varphi \rightarrow D\varphi$
- $D\varphi \rightarrow \neg B\varphi$
- $\neg D\varphi \rightarrow B\neg D\varphi$
- $\neg D\varphi \rightarrow DD\varphi$

- $\neg B\varphi \rightarrow DB\varphi$

As in the cases of *P*-logic and *BP*-logic, the corresponding intuitively incorrect principle, $D\varphi \wedge D\psi \rightarrow D(\varphi \vee \psi)$ can also be avoided in the *BD*-logic. It may be very hard to believe that your friend Craig is the traitor and even that another close friend Denis is the traitor, but circumstantial evidence may make it perfectly plausible that one of them is.

3.3 Preference

There is a very close relationship between an agent's beliefs and her preferences which has been extensively discussed in (de Jongh and Liu (2006), Liu (2008)). Based on the ideas from *optimality theory*, intrinsic preference on the basis of priority sequences $P_1 >> \cdots >> P_n$ is formulated. Here, the $P_i's$ are first-order formulas with exactly one free variable, which is common to all of them. Preferences over objects can be defined in terms of these sequences. The basic idea is to define objective preference by:

$$Pref(d,e) \Leftrightarrow \exists i(P_i d \wedge \neg P_i e) \wedge \forall j < i\,(P_j d \leftrightarrow P_j e)$$

Let us give an example. Alice now has a bunch of applicants for a simple position. She still judges them on a yes-no basis, but now in regard to three aspects: are they strong enough (P_1), can they drive a truck sufficiently well (P_2), do they understand English well enough (P_3). The aspects are ordered in the way described above, i.e., if Jennifer is strong but a poor driver who doesn't speak english, she is graded higher objectively than Karl, an excellent driver with fluent english, but a weakling.
For subjective preferences over objects, which in fact are considered to be influenced by beliefs, several options are considered in the papers mentioned, their meanings are more or less obvious.

$$Pref(d,e) \Leftrightarrow \exists i(B(P_i d) \wedge \neg B(P_i e) \wedge \forall j < i(B(P_j d) \leftrightarrow B(P_j e)))$$

$$Pref(d,e) \Leftrightarrow \exists i(\neg B(\neg P_i d) \wedge B(\neg P_i e) \wedge \forall j < i(B(\neg P_j d) \leftrightarrow B(\neg P_j e)))$$

$$Pref(d,e) \Leftrightarrow \exists i\,((B(P_i d) \wedge \neg B(P_i e)) \vee (\neg B(\neg P_i d) \wedge B(\neg P_i e)) \wedge \forall j < i\,((B(P_j d) \leftrightarrow B(P_j e)) \wedge (B(\neg P_j d) \leftrightarrow B(\neg P_j e))))$$

The first option directly subjectivies the original idea, the criteria are made a matter of belief. But, returning to the example, Alice may not be able to make up her mind about the strength of Malcolm. The second option says then that,

if she judges the driving capabilities of Lars to be clearly insufficient, she rates Malcolm higher than Lars.

It is clear that the above three approaches are different attempts to express that up to a certain level of the priority sequence the degree of belief in the objects d and e having the mentioned properties is the same and that at the next level the degree of belief in d having the right property is greater than that in e having it. Here we can express this directly in the language as below, giving one uniform definition.

$$Pref(d,e) \Leftrightarrow \exists i(P_i d \succ_B P_i e \ \wedge \forall j < i(P_j d \equiv_B P_j e)).$$

As in the introduction, we point out that for many decisions involving preference grading abilities may be unavoidable. Here though, we just look at decisions involving yes-no questions.

4 Safe Belief

The notion of 'safe belief' has been introduced in Baltag and Smets (2008). The authors gave this name to single out those *beliefs* "that are *safe* to hold, in the sense that no future learning of truthful information will force us to revise them". It closely related to "Stalnaker knowledge" Stalnaker (2006), where evidence is considered as true information. The safe belief modality is generally denoted by $\Box$. Evidently, 'safe beliefs' are *truthful* ($\Box\varphi \models \varphi$) and *positively introspective* ($\Box\varphi \models \Box\Box\varphi$), but not necessarily *negatively introspective* (in general, $\neg\Box\varphi \not\models \Box\neg\Box\varphi$).

Adding safe belief to our ordering framework is interesting both from the technical as well as intuitive point of view. This is because in the interpretation of Baltag and Smets (2008) there is a very close relationship between the notion of safe belief and the plausibility ordering.

In the *plausibility models*, the truth definition of $\Box\varphi$ is given by the following clause:

$$\mathcal{M}, s \models \Box\varphi \text{ iff } \mathcal{M}, t \models \varphi \text{ for all worlds } t \leq s.$$

which says that φ can be safely believed at some world s if it holds at all the worlds which are at least as plausible as s. In the following we will introduce the safe belief modality in the setting of *KD45−O*, and give a complete axiomatization of this logic. The language of the logic *KD45−OS* is defined as follows:

Definition 4.1. Given a countable set of atomic propositions Φ, formulas φ are defined inductively:

$$\varphi := \bot \mid p \mid \neg\varphi \mid \varphi \vee \varphi \mid B\varphi \mid \Box\varphi \mid \varphi \geqslant_B \psi$$

where $p \in \Phi$.

We now present the axioms of the logic *KD45−OS* in which the operators U and E are defined as before. Together with the axioms and rules of the *KD45*-logic of beliefs, and the relevant ordering axioms, viz. refl, trans, lin, center, existence, $U \geqslant_B$-axiom and the *S*4-axioms and rules for the safe belief $\Box$ operator, we will have the following extra axioms,

$$(\Box\varphi \wedge \neg\Box\psi) \rightarrow (\varphi >_B \psi) \quad (\Box\text{order-axiom})$$
$$(\varphi \geqslant_B \psi) \rightarrow \Box(\varphi \geqslant_B \psi) \quad (\Box\text{intros-axiom1})$$
$$(\varphi >_B \psi) \rightarrow \Box(\varphi >_B \psi) \quad (\Box\text{intros-axiom2})$$

The $\Box$order axiom generalizes the center axiom. It expresses that, if a set X contains all worlds with a certain grade of plausibility or higher, and Y does not, then $X >_{B_a} Y$. In addition to all these, the following axiom relates the operator $\Box$ with B.

$$\Box\varphi \rightarrow B\varphi \quad (\Box B\text{-axiom})$$

The intros-axioms(1-2) and the un.center axioms of *KD45−O* are derivable from *KD45−OS*. We can also derive:

$$U\varphi \rightarrow \Box\varphi$$

Theorem 5. *The logic KD45−OS is sound and its validities can be completely axiomatized by the following axioms and rules.*

a) all KD45−O axioms and rules

b) S4-axioms and rules for the modal operator $\Box$

c) ordering axioms:

$\varphi \geqslant_B \varphi$ *(refl-axiom)*

$(\varphi \geqslant_B \psi) \wedge (\psi \geqslant_B \chi) \rightarrow \varphi \geqslant_B \chi$ *(trans-axiom)*

$(\varphi \geqslant_B \psi) \vee (\psi >_B \varphi)$ *(lin-axiom)*

$U(\Box\varphi \to \Box\psi) \vee U(\Box\psi \to \Box\varphi)$ *(□lin-axiom)*

$(B\varphi \wedge \neg B\psi) \to (\varphi \succ_B \psi)$ *(center-axiom)*

$(\Box\varphi \wedge \neg\Box\psi) \to (\varphi \succ_B \psi)$ *(□order-axiom)*

$(\varphi \succcurlyeq_B \psi) \to \Box(\varphi \succcurlyeq_B \psi)$ *(□intros-axiom1)*

$(\varphi \succ_B \psi) \to \Box(\varphi \succ_B \psi)$ *(□intros-axiom2)*

$U(\varphi \to \psi) \to (\psi \succcurlyeq_B \varphi)$ *($U \succcurlyeq_B$-axiom)*

$\varphi \to E\varphi$ *(existence axiom)*

d) $\Box\varphi \to B\varphi$ *(□B-axiom)*

e) inclusion rule:

$$\frac{\varphi \to \psi}{\psi \succcurlyeq_B \varphi}$$

Proof. Assume $\nvdash_{KD45-OS} \varphi$. We will have to construct a counter-model to φ which is a $KD45-OS$-model. We take a finite adequate set Φ containing φ. Consider the m.c. (maximally consistent) subsets of Φ. In particular consider such an m.c. set Φ_0 containing $\neg\varphi$.

Define the plausibility ordering among m.c. sets as follows: $P \leq Q$ iff for all $\Box\psi$ in the adequate set, if $\Box\psi$ is in P, then $\Box\psi$ and ψ are in Q. Then immediately we have that $\leq$ is reflexive and transitive.

The relations $\mathcal{R}_B$ and $\mathcal{R}_U$ are defined as follows:

$$\begin{array}{lll} P\mathcal{R}_B Q & \text{iff} & \text{(1) for all } B\varphi \text{ in } P,\ \varphi \text{ as well as } B\varphi \text{ are in } Q, \\ & & \text{(2) for all } \neg B\varphi \text{ in } P,\ \neg B\varphi \text{ in } Q. \\ P\mathcal{R}_U Q & \text{iff} & \text{(1) for all } U\varphi \text{ in } P,\ \varphi \text{ as well as } U\varphi \text{ are in } Q, \\ & & \text{(2) for all } \neg U\varphi \text{ in } P,\ \neg U\varphi \text{ in } Q \end{array}$$

As in the proof of Theorem 2.3, we can show that $\mathcal{R}_U$ will be an equivalence relation and $\mathcal{R}_B$ a euclidean subrelation of $\mathcal{R}_U$.

It follows from □intros-axioms that $U\varphi \to \Box\varphi$ is derivable, and so $\leq$ is a sub-relation of $\mathcal{R}_U$. From the axioms relating $\Box$ and B, it follows that $\mathcal{R}_B$ is a sub-relation of $\leq$.

We now take the submodel generated by $\mathcal{R}_U$ from Φ_0. The set of worlds W of our model will be the set of worlds in this submodel and the $\mathcal{R}_B$ and $\mathcal{R}_U$ the

restrictions of the original $\mathcal{R}_B$ and $\mathcal{R}_U$ to this submodel. $\mathcal{R}_U$ is now the universal relation. As before, we write $\mathcal{B}$ for the set of $\mathcal{R}_B$-reflexive elements. Because of the $\Box$lin-axiom $\leq$ becomes linear in this model. So, with respect to the modal operators B and E and $\Box$ the model behaves properly. We will now have to order $\mathcal{P}(W)$ in the proper way, which can be done as in the proof of Theorem 2.3, using the $\Box$ order axiom in addition to the center axiom. $\Box$

We should mention here that, according to Baltag and Smets (2008), belief and conditional belief can be expressed in terms of knowledge and safe belief as,

$$B^{\psi}\varphi := \widehat{K}\psi \rightarrow \widehat{K}(\psi \wedge \Box(\psi \rightarrow \varphi)),$$
$$B\varphi := B^{\top}\varphi,$$

where, $\widehat{K}\psi := \neg K \neg \psi$. They gave complete axiomatizations for conditional doxastic logic (logic of conditional belief) as well as the logic of knowledge and safe beliefs. We do not consider knowledge but for this part of the discussion it can be replaced by U. Neither do we talk about conditional belief here, but belief can be defined in terms of the existential modality and safe belief (and therefore, in terms of safe belief and belief ordering) as follows:

$$B\varphi := E\Box\varphi$$

Once we have in this manner the modal operator B as a defined concept, we can easily derive all its well-known properties in $KD45{-}OS$, but if that holds fully for its relations with $\succcurlyeq_B$ remains to be seen.

5 Multi-agent system

The main focus of this paper has been on beliefs and strengths of beliefs of a single agent. The whole idea can be generalized to the multi-agent framework which is what we do in the following. The language of the logic of belief ordering in the multi-agent case, $KD45{-}O_M$ can be defined as follows:

Definition 5.1. Given a finite set of agents A, and a countable set of atomic propositions Φ, formulas φ are defined inductively:

$$\varphi := \bot \mid p \mid \neg\varphi \mid \varphi \vee \varphi \mid B_a\varphi \mid \varphi \succcurlyeq_{B_a} \psi$$

where $p \in \Phi$, $a \in A$.

The indices in the belief modality and in the ordering formula denote the agents whose beliefs or strengths of beliefs are considered. The operators $\succ_{B_a}$ and U_a are defined in the usual way. The fact that U is also indexed may surprise the reader for a moment but it is the only coherent way to extend the one agent case. Existence of a location for a proposition to be true meant for us that for the one agent case, belief in the proposition was stronger than belief in a contradiction. With more agents, we may have those who differ in regard to the existence of propositions: more worlds will have to be added to the model, and it will not stop there: there is no reason for E_aE_b to be equivalent to E_a or E_b, etc.

Keeping all these considerations in mind, the models for $KD45{-}O_M$ have to be suitable multi-agent generalizations of those for $KD45{-}O$. The basic idea to consider here is that we can no longer rule out worlds that are *impossible* for an agent a. They might well be possible for another agent b and also have to be considered while talking about agent a's belief about agent b's beliefs and so on. Evidently, the earlier *plausibility ordering* and *set ordering* of worlds will get indexed by agents (one for each agent), and the global concept of belief will give way to more local concepts of beliefs. This fact becomes apparent in the syntax also, with the introduction of formulas like $U_a\varphi$. The notion of *comparative classes* Baltag and Smets (2008) which gives the set of worlds that an agent considers relevant while positioned at her current world comes into play. Formally, a comparative class of some world is just the set of worlds that are related to the current world by the plausibility order. To give meaning to agents' beliefs, strength of beliefs, these relevant worlds are needed to be considered only, unlike the single agent case, where the whole model is taken into account.

Definition 5.2. Given a finite set of agents A, a $KD45{-}O_M$ model is defined to be a structure $\mathcal{M} = (S, \{\leq_a : a \in A\}, \{\geq_{B_a} : a \in A\}, V)$, where S is a non-empty finite set of states, V is a valuation assigning truth values to atomic propositions in states, and for each a, $\leq_a$ is a pre-order relation over S, which forms a partition of S given by $\sim_a = \leq_a \cup \geq_a$, an equivalence relation over S. Finally, for each $a \in A$, $\geq_{B_a}$ is a quasi-linear order relation over $\mathcal{P}(T)$ for each equivalence class T of $\sim_a$, satisfying the conditions:

(1) If $X \subseteq Y \subseteq T$, then $Y \geq_{B_a} X$

(2) if $\mathcal{B}_a \subseteq T$ is the set of a-plausible worlds, truth on which suffices to make an assertion to be believed (that is, the set of all $\leq_a$-minimal worlds in

T), then $\mathcal{B}_a \subseteq X \subseteq T \wedge \mathcal{B}_a \nsubseteq Y \subseteq T \Rightarrow X >_{B_a} Y$, where $>_{B_a}$ denotes the corresponding strict ordering.

(3) If $X \subseteq T$ is non-empty, then $X >_{B_a} \emptyset$.

For any $s \in S$, let s_a denote the set of all members of S which are $\sim_a$-equivalent to s. The truth definition for formulas φ in a $KD45$–O_M model $\mathcal{M}$ is as usual with the following clauses for the belief and ordering modalities.

$\mathcal{M}, s \models B_a\varphi$ iff $\mathcal{M}, t \models \varphi$ for all $\leq_a$-minimal worlds $t \in s_a$.

$\mathcal{M}, s \models \varphi \succcurlyeq_{B_a} \psi$ iff $\{t \in s_a \mid \mathcal{M}, t \models \varphi\} \geq_{B_a} \{t \in s_a \mid \mathcal{M}, t \models \psi\}$.

We considered $\succcurlyeq_B$ to be a global notion – if $\varphi \succcurlyeq_B \psi$ is true anywhere in the model, it is true everywhere. But in the multi-agent case, $\succcurlyeq_{B_a}$ does become to a certain extent state-dependent, which is intuitive as different agents may perceive the world in different ways. But, of course, the notion does stay a global notion within each $\sim_a$ equivalence class. From the definition of $\succ_{B_a}$, it follows that,

$\mathcal{M}, s \models \varphi \succ_{B_a} \psi$ iff $\{t \in s_a, \mid \mathcal{M}, t \models \varphi\} >_{B_a} \{t \in s_a \mid \mathcal{M}, t \models \psi\}$.

Thus, $\succ_{B_a}$ also becomes a more local notion. We will now define the corresponding localized universal modality U_a for each agent $a \in A$. As earlier, the modality $E_a\varphi$ (the abbreviated form of $\neg U_a \neg\varphi$) can be defined as $\varphi \succ_{B_a} \bot$, and hence $U_a\varphi$ as $\bot \succcurlyeq_{B_a} \neg\varphi$. The formula $U_a\varphi$ expresses that φ is true in all $a-$*accessible* worlds in the model, whereas $E_a\varphi$ stands for existence of a possible $a-$*accessible* world in the model where φ is true. Evidently, we have,

$\mathcal{M}, s \models U_a\varphi$ iff $\mathcal{M}, t \models \varphi$ for all worlds $t \in s_a$.

As earlier, each of these U_a modalities needs to satisfy the $S5$-axioms that hold for U Goranko and Passy (1992) plus the axiom $B_a\varphi \rightarrow U_aB_a\varphi$, which expresses that B_a is a global notion in each of the $\sim_a$ equivalence classes, where U_a expresses this universality.

The logic $KD45 - O_M$ arises from the logic $KD45 - O$ by indexing, for each agent a, in each axiom both the operator B and $\succcurlyeq_B$ by a so that, for each agent the same axioms arise with B_a instead of B and $\succcurlyeq_{B_a}$ instead of $\succcurlyeq_B$.

Definition 5.3. The system $KD45 - O_M$ consists of, for each agent $a \in A$,

a) all $KD45$ axioms and rules for B_a

b) ordering axioms:

$\varphi \geqslant_{B_a} \varphi$ (refl-axiom)

$(\varphi \geqslant_{B_a} \psi) \wedge (\psi \geqslant_{B_a} \chi) \rightarrow \varphi \geqslant_{B_a} \chi$ (trans-axiom)

$(\varphi \geqslant_{B_a} \psi) \vee (\psi >_{B_a} \varphi)$ (lin-axiom)

$(B\varphi \wedge \neg B\psi) \rightarrow (\varphi >_{B_a} \psi)$ (center-axiom)

$(\varphi \geqslant_{B_a} \psi) \rightarrow B(\varphi \geqslant_{B_a} \psi)$ (intros-axiom1)

$(\varphi >_{B_a} \psi) \rightarrow B(\varphi >_{B_a} \psi)$ (intros-axiom2)

$\bot \geqslant_{B_a} \neg(\varphi \rightarrow \psi) \rightarrow (\psi \geqslant_{B_a} \varphi)$ ($U \geqslant_{B_a}$-axiom)

$\varphi \rightarrow (\varphi >_{B_a} \bot)$ (existence axiom)

$(B\varphi >_{B_a} \bot) \rightarrow B\varphi$ (un.center-axiom)

c) inclusion rule:

$$\frac{\varphi \rightarrow \psi}{\psi \geqslant_{B_a} \varphi}$$ (inclusion rule)

As in the single-agent case we have the following result.

Theorem 6. *$KD45{-}O_M$ is sound and complete with respect to $KD45{-}O_M$ models.*

The completeness proof is a generalization of the completeness proof for $KD45 - O$ by executing within each U_a-equivalence class the same prodeure as in that proof. We refrain from going into the proof details. Evidently, $KD45 - O_M$ is also decidable.

6 World ordering versus set ordering: a discussion

Why have a set of worlds ordering when one has a world ordering available? If one wants to define $\varphi >_B \psi$ and $\varphi \geqslant_B \psi$ in terms of the plausibility ordering of the worlds then the following option comes to mind: interpret $\varphi \geqslant_B \psi$ as, for each ψ-world there exist φ-worlds which are at least as plausible (similar to the proposal in Lewis (1973)). If one does this however, $B\varphi$ becomes equivalent to $\varphi \geqslant_B \neg\bot$. Also, contrary to our aims, no distinction in strength of belief can be made between propositions which are believed. More sophisticated reductions of strengths of beliefs to the plausibility ordering of the worlds will

have undesirable consequences as well. This becomes very clear in the Section 3.1. Restricting our ordering to formulas $\varphi \succ_B \neg\varphi$ gives rise to a plausibility logic which seems to validate exactly the formulas that we want, and that logic is non-normal but monotonic. The standard semantics for such logics is neighborhood semantics which uses sets, and some semantics involving sets of worlds seems necessary.

To do away with the whole issue we introduced a *set-plausibility* ordering $\geq_B$ between sets of worlds and put very minimal requirements on this ordering.

If one keeps the world ordering and the set ordering independent from each other, then there are no real difficulties in adding dynamics to the system. The starting point for that lies in Subsection 4. The dynamics of safe belief have been well-established Baltag and Smets (2008), and, using the results of this subsection, a dynamic version seems well within reach.

If, alternatively, one wants to define the world ordering in terms of the set ordering, one may add a principle introduced in Subsection 2.2, let us call it center-axiom2,

$$(B\neg\psi \wedge \neg B\neg\varphi) \rightarrow (\varphi \succ_B \psi).$$

This expresses that, if Y is disjoint from the center $\mathcal{B}$ and X intersects the center, then $X >_B Y$. In the completeness proof model each singleton set $\{s\}$ is uniquely determined by a formula φ_s. If we define the plausibility order by $s \leq t$ iff $\varphi_s \geq_B \varphi_t$, then this ordering gets the right properties.

There is a catch however in proceeding this way, there is no reason for all the worlds in the center to get the same maximal degree of plausibility. For our intuitions this is no great problem, but it does mean a definite obstacle in making our system dynamic, since in the standard plausibility models the center consists of the most plausible worlds, and this property is used to single out the new center after e.g. a public announcement has been received. We do have ideas to solve this problem, but that is for a future occasion.

7 Conclusion and further work

An explicit ordering of formulas to compare the strengths of belief is introduced. A complete axiomatization for this belief logic with explicit ordering is provided with respect to a semantics that includes a set ordering in addition to the standard plausibility ordering. The notion aids in giving intuitive formulations for various related concepts like universality as well as some other

epistemic attitudes - much older and thoroughly discussed notions like *universality* and *preference*, together with relatively newer ones like *plausibility* and *disbelief*. Independent axiomatizations for the logics of plausibility, belief and plausibility as well as belief and disbelief are also provided. Interplay of belief ordering with the concept of safe beliefs is discussed. Lastly, we lift the proposed framework to a multi-agent setting.

In Section 6 we discussed the possibilities and problems connected with providing a dynamic version of the present work. This seems definitely promising.

Acknowledgements We thank the anonymous referees of FAMAS 2009 for their close reading and extensive comments, which helped us to improve the paper. We also thank Alexandru Baltag and Sonja Smets for presenting our paper at FAMAS 2009, and their subsequent comments which inspired us in writing this version. The second author thanks the Center for Soft Computing Research, Indian Statistical Institute, Kolkata for the congenial atmosphere she had there while doing this work during February, 2008 to February, 2009 and also August, 2009. She also acknowledges NWO grant # 600.065.120.08N201.

References

A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *Logic and the Foundations of Game and Decision Theory, Texts in Logic and Games*, volume 3, pages 9–58. Amsterdam University Press, 2008.

O. Board. Dynamic inetractive epistemology. *Games and Economic Behavior*, 49: 49–80, 2004.

J. Burgess. Probability logic. *Journal of Symbolic Logic*, 34(2):264–274, 1969.

M. Chakraborty and S. Ghosh. Belief-disbelief interface: A bi-logical approach. *Fundamenta Informaticae*, to appear.

B. Chellas. *Modal logic: an introduction*. C.U.P., 1980.

S. Chopra, J. Heidema, and T. Meyer. Logics of belief and disbelief. In *Proceedings of the ninth International Workshop on Non-Monotonic Reasoning*, 2002.

D. de Jongh and F. Liu. Optimality, belief and preference. In S. Artemov and R. Parikh, editors, *Proceedings of the Workshop on Rationality and Knowledge, ESSLLI*, 2006.

N. Friedman and J. Halpern. Plausibility measures and default reasoning. *Journal of the ACM*, 48(4):648–685, 2001.

P. Gardenfors. Qualitative probability as an intentional logic. *Journal of Philosophical Logic*, 4:171–185, 1975.

P. Gärdenfors and D. Makinson. Revisions of knowledge systems and epistemic entrenchment. In M. Vardi, editor, *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 83–95. Los Altos: Morgan Kaufmann, 1988.

J. Gerbrandy. *Bisimulation on Planet Kripke*. PhD thesis, University of Amsterdam, 1999.

A. Ghose and R. Goebel. Belief states as default theories: Studies in non-prioritised belief change. In H. Prade, editor, *Proceedings of the 13th European Conference on Artificial Intelligence*, pages 8–12, 1998.

A. Gomolinska. On the logic of acceptance and rejection. *Studia Logica*, 60:233 – 251, 1998.

A. Gomolinska and D. Pearce. Disbelief change. *Electronic essays on the occasion of the fiftieth birthday of Peter Gardenfors*, 2001.

V. Goranko and S. Passy. Using the universal modality: Gains and questions. *Journal of Logic and Computation*, 2(1):5–30, 1992.

J. Halpern. Defining relative likelihood in partially-ordered preferential structures. *Journal of AI Research*, 7:1–24, 1997.

J. Halpern. *Reasoning About Uncertainty*. MIT Press, 2003.

J. Halpern and Y. Moses. A guide to completeness and complexity for modal logics of knowledge and belief. *Artificial Intelligence*, 54:319–379, 1992.

H. Hansen. Monotonic modal logics. In *ILLC-prepublication series*, PP-2003-24.

S. Kripke. Semantical considerations on modal logics. *Acta Philosophica Fennica*, 16:83–94, 1963.

D. Lewis. *Counterfactuals*. Blackwell and Harvard U.P., 1973.

F. Liu. *Changing for the better: Preference Dynamics and Agent Diversity*. PhD thesis, University of Amsterdam, 2008.

R. Moore. Semantical considerations on nonmonotonic logic. *Artificial Intelligence*, 25:75–94, 1985.

K. Segerberg. Qualitative probability in a modal setting. In J. Fenstad, editor, *Proceedings of the 2nd Scandinavian Logic Symposium*, Amsterdam, 1971. North-Holland.

W. Spohn. Ordinal conditional functions. a dynamic theory of epistemic states. In W. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change, and Statistics*, volume II, pages 105–134. Kluwer, Dordrecht, 1988.

R. Stalnaker. On logics of knowledge and belief. *Philosophical Studies*, 128(1): 169–199, 2006.

J. van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logic*, 17(2):129–155, 2007.

Epistemic Closure and Epistemic Logic I

Wesley H. Holliday

Department of Philosophy, Stanford University
weshlliday@stanford.edu

Abstract

This paper develops a formal framework to study epistemic closure, using epistemic-logical models of the *relevant alternatives*, *tracking*, and *safety* theories of knowledge. The main result is a complete characterization of the epistemic closure principles that hold according to these theories, as formalized. Analysis of this Closure Theorem shows that two parameters of a modal theory of knowledge affect whether the theory preserves closure.[1]

1 Introduction

At its simplest, the claim that *knowledge is closed under known implication* is the claim that if one knows that φ and knows that φ implies ψ, then one knows that ψ: $(K\varphi \wedge K(\varphi \to \psi)) \to K\psi$, in the language of epistemic logic. Although few have objected to the validity of $(\varphi \wedge (\varphi \to \psi)) \to \psi$, many have objected to the closure of knowledge under known implication.

According to one form of objection, common in epistemic logic, the claim is not true in general, because an agent with bounded rationality may fail to draw the inference to ψ, may forget something in the process, etc. According to another form of objection, made famous in epistemology by Dretske (1970) and Nozick (1981), the claim is not true even for agents with ideal rationality,

[1] This paper is an abridged preprint of the first of a series of three papers on epistemic closure and epistemic logic.

who draw all the valid inferences and forget nothing. The subsequent debate over epistemic closure principles has been called "one of the most significant disputes in epistemology over the last forty years or so" (Kvanvig 2006, p. 256).

The closure of knowledge under known implication (hereafter referred to as '**K**', the name of the modal axiom given above) is one closure principle among (infinitely) many. Although Dretske (1970) denied **K**, he accepted other closure principles, such as closure under conjunction elimination ($K(\varphi \wedge \psi) \rightarrow K\varphi$) and closure under disjunction introduction ($K\varphi \rightarrow K(\varphi \vee \psi)$) (p. 1009). Nozick (1981) was prepared to give up closure under conjunction elimination (p. 228), although not closure under disjunction introduction (p. 692, n. 64).

A general argument strategy against those who deny **K** is to show that they are committed to giving up other attractive closure principles. In short, closure failures spread, and they spread to where no one wants them. Pressing such a Problem of Containment has an advantage over other strategies, for it appeals to principles that both sides of the debate over **K** are likely to accept, rather than merely insisting on the validity of **K**.[2] What is missing is a systematic assessment of the extent of this problem for standard theories of knowledge.

In this paper, I perform such an assessment for a family of related theories. In particular, I introduce epistemic-logical models of the *relevant alternatives* (RA) theories of Lewis (1996) and Heller (1989, 1999), the *tracking* theory of Nozick (1981), and the *safety* theory of Sosa (1999). The main result is a complete characterization of the closure principles that hold according to these theories, as formalized. I will briefly preview the consequences of this Closure Theorem.

First, except for Lewis's theory, the other RA, tracking, and safety theories cited suffer from widespread closure failures, far beyond the failure of **K**, which few if any proponents of these theories are willing to accept. (Surprisingly, the closure principles that fail on these different theories are *exactly the same*.)

Second, while closure failures go too far on all of the theories but Lewis's, in another way they do not go far enough for the purposes of some proponents of the theories. Closure principles that appear just as dangerous as **K** in arguments for *skepticism* hold for these theories, which defeats the purpose of invoking the failure of **K** in response to skepticism, as Nozick and Dretske famously do.

Third, analysis of the Closure Theorem shows that two parameters of a modal theory of knowledge affect whether the theory preserves closure. Each parameter has two values, generating four possible parameter settings, with respect to which each theory can be classified. Only Lewis's theory, with its

[2]Hawthorne (2004, p. 41) pursues a version of this strategy, which I will discuss in a sequel to this paper. According to Cohen (2002, p. 312), so did Kripke in unpublished lectures. Lawlor (2005, p. 44) makes the methodological point about the advantage of this strategy.

unique parameter setting, preserves closure. This presents a dilemma. In the terminology of Dretske, the knowledge operator in Lewis's theory is *fully penetrating*. For all of the other theories, it is not even *semi-penetrating*, according to Dretske's characterization. Finding a theory of knowledge between these extremes seems to require abandoning the "world-ordering" picture employed by the standard theories. (In a sequel to this paper, we do just that.)

The common feature of the theories we will study is some counterfactual or counterfactual-like condition on knowledge, relating what an agent knows to what holds in *close counterfactual possibilities* or *relevant epistemic possibilities*. Vogel (2007) characterizes *subjunctivism* as "the doctrine that what is distinctive about knowledge is essentially modal in character, and thus is captured by certain subjunctive conditionals" (p. 73), and some versions of the RA theory have a similar flavor.[3] Reflecting this common feature, our formal framework is based on the standard semantics for subjunctive conditionals in the style of Lewis 1973 and Stalnaker 1968. As a consequence, the epistemic logics studied here behave very differently than epistemic logics in the style of Hintikka 1962.

In §2, I introduce our running example, which motivates questions of epistemic closure. §3 develops the formal framework for the study of closure in RA (§3.1) and subjunctivist (§3.2) theories, culminating in the Closure Theorem (§3.3) and the analysis of theory parameters and closure failures (§3.4, Table 1). Finally, in §4 I conclude with a summary of the topics treated in the full version of this abridged preprint, as well as in the papers to follow in this series, and I offer some brief reflections on the methodology of our epistemic-logical approach. An Appendix follows with proofs of the main results.

2 Background

Example 1. Two medical students, A and B, are subjected to a test. Their professor introduces them to the same patient, who presents various symptoms, and the students are to make a diagnosis of the patient's condition. After some independent investigation, both students conclude that the patient has a common condition c. In fact, they are both correct. Yet only student A passes the test. For the professor wished to see if the students would check for another common condition c', which causes the same visible symptoms as c. While student A ran laboratory tests to rule out c' before making the diagnosis of c, student B made the diagnosis of c after only a physical exam.

[3] To be careful, the view that knowledge is modal and the view that it is captured by subjunctive conditionals are different. For example, Lewis (1996) adopts the first but not the second.

In evaluating the students, the professor concludes that although both gave the correct diagnosis of c, student B did not know that the patient's condition was c, since he did not rule out the alternative of c'. Had the patient's condition been c', student B might still have made the diagnosis of c, since the physical exam would not have revealed a difference. Student B was *lucky*. The condition he associated with the patient's visible symptoms happened to be the condition the patient had, but if the professor had chosen a patient with c', student B might have made a misdiagnosis. By contrast, student A secured against this possibility of error by running the lab tests. For this reason, the professor judges that student A knew the patient's condition and passed the test.

Of course, student A did not secure against *every* possibility of error. Suppose there is an extremely rare disease[4] x such that people with disease x appear to have c on many lab tests, even though people with x are *immune* to c, and only extensive further testing can detect the presence of x in its early stages. Should we say that student A did not know that the patient's condition was c after all, since she did not rule out the possibility of x? The requirement that one rule out *all* possibilities of error seems to make knowledge impossible, since there are always some possibilities of error—however remote and far-fetched—that are not eliminated by one's evidence and experience. However, if no one had any reason to think that the patient may have had the rare disease x, then it should not have been necessary to rule out such a remote possibility in order to know that the patient has some common condition.[5]

If one accepts the foregoing reasoning, then one is close to a denial of closure under known implication (**K**). For suppose student A knows that people with c do not have x (because x confers immunity to c), which we will write as

(1) $K(c \to \neg x)$.

Since student A did not run any tests that could possibly detect the presence or absence of x, she does not know that the patient does not have x:

(2) $\neg K \neg x$.

Then given the judgment that A knows that the patient has condition c,

(3) Kc,

[4]Perhaps it has never been documented, but it is a possibility raised by a hypochondriac.

[5]Local skeptics about medical knowledge may substitute one of the standard cases in the literature with a similar structure involving, e.g., zoo animals, red surfaces, or BIVs.

we have a clear violation of the following instance of **K**:

(4) $(Kc \wedge K(c \rightarrow \neg x)) \rightarrow K\neg x$.

To retain **K**, one must say either that A does not know the patient's condition after all or that one can know that a patient does not have a rare disease without running any of the diagnostic tests for the disease.[6] The first option leads to radical skepticism, since for any condition c, one can always consider the possibility of some x, however far-fetched, related to c as in our example. The second leads to a problematic kind of "easy knowledge" (Cohen 2002).

Dretske (1970) and Nozick (1981) take the likes of (1)-(3), a version of the now standard "skeptical paradox" (Cohen 1988, DeRose 1995), to show that knowledge fails to satisfy **K**. This failure has nothing to do with the finite reasoning capacities, memory, etc., of agents. According to Dretske (1970), **K** would fail even for "ideally astute logicians," who are "fully appraised of all the necessary consequences . . . of every proposition" (p. 1010). We will cash out this description as follows: first, one knows all valid logical principles (*validity omniscience*);[7] second, one believes all the logical consequences of the (set of) propositions one believes (*full doxastic closure*). Dretske's explanation for why **K** fails even for such ideal logicians is in terms of the RA theory. (We discuss Nozick's view in §3.2.) To know that p is to (have a true belief that p and) to have *ruled out the relevant alternatives to p*. In coming to know c and $c \rightarrow \neg x$, the agent rules out certain relevant alternatives. In order to know $\neg x$, the agent must also rule out certain relevant alternatives. But the relevant alternatives in the two cases *are not the same*. We have already argued that x is not a relevant alternative that must be ruled out in order for Kc to hold. But x certainly is a relevant alternative that must be ruled out in order for $K\neg x$ to hold (cf. Remark 3.3 of §3.1). It is because the relevant alternatives may be different for the antecedent and the consequent that **K** does not hold in general.

In an influential objection to Dretske, Stine (1976) argued that to allow for the relevant alternatives to be different for the premises and conclusion of an inference "would be to commit some logical sin akin to equivocation" (p. 256). Yet as Heller (1999) points out in Dretske's defence, a similar charge of equivocation could be made (incorrectly) against accepted counterexamples to the principles of transitivity or antecedent strengthening for counterfactuals. If

[6] This statement of the dilemma ignores the option of *contextualism* , which we will study in a successor of this paper. Stine (1976), Lewis (1996), and Cohen (1988) propose contextualist versions of the RA theory, while DeRose (1995) proposes a contextualist version of Nozick's tracking theory.

[7] A stronger property of *consequence omniscience*, that one knows all the logical consequences of what one knows, implies validity omniscience, but not vice versa.

we take a counterfactual $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ to be true just in case the "closest" φ-worlds are ψ-worlds, then the inference from $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ to $(\varphi \wedge \chi) \mathrel{\Box\!\!\rightarrow} \psi$ fails because the closest $(\varphi \wedge \chi)$-worlds may not be the same as the closest φ-worlds. Heller argues that there is no equivocation in such counterexamples since we use the same, fixed similarity ordering of worlds to evaluate the different conditionals. Similarly, in the example of closure failure, the most relevant $\neg c$-worlds may be distinct from the most relevant x-worlds (so one can rule out the former without ruling out the latter), even assuming a fixed relevance ordering over the set of worlds. In this defense of Dretske, Heller brings the RA theory closer to the subjunctivist theories that place counterfactual conditions on knowledge.

3 A Formal Study of Closure Failure

In this section, we undertake our formal study of closure failure. Throughout, we use the language of propositional epistemic logic, generated from atomic sentences $p, q, r, \ldots$ using connectives $\neg$ and $\wedge$ (from which $\vee$, $\rightarrow$, and $\leftrightarrow$ are defined) and a knowledge operator K. We write $\varphi, \psi, \alpha, \beta$, etc., for arbitary formulas of the language. A formula is *propositional* iff it does not contain K.

Definition 3.1 (Closure Principle)**.** A *closure principle* is any formula of the form $(K\varphi_1 \wedge \cdots \wedge K\varphi_n) \rightarrow K\psi$, where $\varphi_1, \ldots, \varphi_n$ and ψ are propositional formulas and $(\varphi_1 \wedge \cdots \wedge \varphi_n) \rightarrow \psi$ is a tautology.

From now on, we will omit parentheses and write $K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi$, with the usual understanding that $\wedge$ and $\vee$ bind more strongly than $\rightarrow$.

Examples of closure principles include $K\alpha \wedge K(\alpha \rightarrow \beta) \rightarrow K\beta$ (closure under known implication); $K\alpha \wedge K\beta \rightarrow K(\alpha \wedge \beta)$ (closure under conjunction introduction); $K(\alpha \wedge \beta) \rightarrow K\alpha$ (closure under conjunction elimination); and $K\alpha \rightarrow K(\alpha \vee \beta)$ (closure under disjunction introduction). Our aim is to investigate which closure principles, among these and others, are valid given different semantics for the knowledge operator K. In doing so, we will gauge the severity of the Problem of Containment, introduced in §1, for theories with closure failure.

3.1 Relevant Alternatives

An important distinction between different versions of the RA theory has to do with logical structure. Dretske (1981) introduces the following definition in developing his version of the theory: "let us call the set of possible alternatives that a person must be in an evidential position to exclude (when he knows

P) the *Relevancy Set*" (p. 371). It is clear from Dretske's definition and the discussion that follows that the choice of the relevancy set depends on P. By contrast, Heller (1999) considers (and rejects) an interpretation of the RA theory according to which "there is a certain set of worlds selected as relevant, and S must be able to rule out the not-p worlds within that set" (p. 197). In this case, the choice of the set of relevant worlds does not depend on P.

The logical distinction is that of $\forall\exists$ vs. $\exists\forall$, which we will mark with two versions of the RA theory. Let W be a set of possible worlds. Following Lewis (1973, p. 46), we take a proposition P to be a subset of W, so that the complement of P in W, $W \setminus P = \{w \in W \mid w \notin P\}$, is the proposition not-P.

- According to an $\text{RA}_{\forall\exists}$ theory, (for every context C and) for every ($\forall$) proposition P, there is ($\exists$) a set of *relevant not-P worlds*, $r_c(P) \subseteq W \setminus P$, such that in order to know P one must rule out the worlds in $r_c(P)$.
- According to an $\text{RA}_{\exists\forall}$ theory, (for every context C) there is ($\exists$) a set of *relevant worlds*, R_C, such that for every ($\forall$) proposition P, in order to know P one must rule out the not-P worlds in that set, i.e., $R_C \cap (W \setminus P)$.

Although this distinction does not appear explicitly in the literature, Dretske (1981) assumes $\text{RA}_{\forall\exists}$, while Lewis (1996) assumes $\text{RA}_{\exists\forall}$. This difference turns out to be at the heart of the disagreement about epistemic closure principles.

Remark 3.1. In our characterization of $\text{RA}_{\forall\exists}$ vs. $\text{RA}_{\exists\forall}$ theories, the parenthetical reference to a context C is important. In a *contextualist* $\text{RA}_{\exists\forall}$ theory, such as Lewis's theory, the set of relevant worlds may change as context changes. Still, for any given context C, there is a set R_C of relevant worlds, which does not depend on the particular proposition in question—unlike in $\text{RA}_{\forall\exists}$ theories, which allow such dependence. The point is that the $\text{RA}_{\forall\exists}$ vs. $\text{RA}_{\exists\forall}$ distinction is about how different theories view the relevant alternatives *with respect to a fixed context*. In this paper, we are interested in which closure principles hold, according to different theories, with respect to a fixed context. In a successor to this paper, we will extend the framework to study context change.

Another general distinction between versions of the RA theory has to do with different notions of ruling out alternatives or eliminating possibilities.

- Lewis (1996) proposes that "a possibility . . . [v] . . . is *uneliminated* iff the subject's perceptual experience and memory in . . . [v] . . . exactly match his perceptual experience and memory in actuality" (p. 553).
- Heller (1999) proposes that "S's ability to rule out not-p be understood thus: S does not believe p in any of the relevant not-p worlds" (p. 198).

In this section, we model the RA theory with Lewis's notion of elimination. In §3.2, we turn to Heller's notion of elimination, which is closely related to Nozick's (1981) tracking theory of knowledge.

To prepare for Definitions 3.2 and 3.3, consider the following RA picture from Heller 1989: "The picture we get is of spheres of possible worlds surrounding the actual world ordered according to how realistic they are, so that those worlds that are more realistic are closer to the actual world than the less realistic ones. To evaluate the claim that S can discriminate between the relevant worlds we examine every not-p world that is realistic enough" (p. 25).

Remark 3.2. Given a set of worlds W, we associate with each $w \in W$ a *relevance relation* $\preceq_w$ that orders worlds in W by how relevant/realistic they are at w. Technically, $\preceq_w$ will be a *preorder* on W that is *total*, *converse well-founded*, and has w as a *maximal* element.[8] A preorder is a reflexive and transitive binary relation. The preorder is total on W iff any two $u, v \in W$ are comparable in relevance at w, i.e., $u \preceq_w v$ or $v \preceq_w u$. The preorder is converse well-founded on W iff for every non-empty subset $S \subseteq W$, there is a world that is *maximally* relevant in S according to $\preceq_w$, i.e., a $v \in S$ such that for all $u \in S$, $u \preceq_w v$. In this sense, we assume that w is maximally relevant in W according to $\preceq_w$.

The picture given by Remark 3.2 leads to the definition of our first class of models. Note that these models represent the epistemic situation of an agent from a third-person perspective. We do not assume that the set of worlds, the relevance orderings, etc., is something that the agent herself has in mind.

Definition 3.2 (RA Model)**.** A *relevant alternatives model* is a tuple $\mathcal{M} = \langle W, \sim, \preceq, V \rangle$ where W is a non-empty set; $\sim$ is an equivalence relation on W; $\preceq$ is a *set* containing, for each $w \in W$, a total and converse well-founded preorder $\preceq_w$ on W in which w is maximal; and V is a valuation function that assigns to each atomic sentence p a set $V(p) \subseteq W$.

We refer to elements of W as "worlds" or "possibilities" interchangeably. As usual, we think of V as encoding the *atomic facts* that hold at each world, by mapping each atom p to the set of worlds $V(p) \subseteq W$ where it holds.

We interpret $w \sim v$ to mean that possibilities w and v are *indistinguishable* for the agent, in Lewis's sense that the agent has the same perceptual experience and memory in w and v.[9] We will also say that v is *uneliminated* at w when $w \sim v$. When $w \nsim v$, we say that v is *eliminated* or *ruled out* at w.

[8] Such a set of preorders is essentially equivalent to one of Lewis's (1973) comparative similarity systems (p. 48) with weak centering (p. 29) and the Limit Assumption (p. 19).

[9] Following Lewis's idea of *exactly matching* experience and memory, Definition 3.2 states that $\sim$ is an equivalence relation. However, we use only the reflexivity of $\sim$ for the results of this paper.

We interpret $u \preceq_w v$ to mean that possibility v is *at least* as relevant at w (hereafter "relevant$_w$") as possibility u. The abbreviation $u \prec_w v$, or equivalently $v \succ_w u$, defined as $u \preceq_w v$ & $v \not\preceq_w u$, indicates that v is *more* relevant$_w$ than u; the abbreviation $u \simeq_w v$, defined as $u \preceq_w v$ & $v \preceq_w u$, indicates that u and v are *equally* relevant$_w$. For notation, we write $\mathrm{Max}_{\preceq_w}(S) = \{v \in S \mid u \preceq_w v \text{ for all } u \in S\}$ for the set of maximally relevant$_w$ possibilities out of a given set of possibilities $S \subseteq W$. We consider any world in $\mathrm{Max}_{\preceq_w}(W)$ to be simply *relevant$_w$*. Finally, the assumption that for all worlds w, w is maximal in $\preceq_w$, amounts to the assumption that for all worlds w, w is relevant$_w$. Lewis (1996) calls this the *Rule of Actuality*, that "actuality is always a relevant alternative" (p. 554).[10]

We now interpret our formal epistemic language in RA models, considering three semantics for knowledge formulas $K\varphi$. We refer to these as C-semantics, for **C**artesian, D-semantics, for **D**retske, and L-semantics, for **L**ewis, respectively. C-semantics is not meant to capture Descartes' view of knowledge. Rather, it is supposed to reflect a high standard for the truth of knowledge claims—knowledge requires ruling out all possibilities of error, however remote—in the spirit of Descartes' worries about error in the First Meditation. D-semantics is *one* way of understanding Dretske's (1981) $\mathrm{RA}_{\forall\exists}$ theory, using Heller's (1989, 1999) picture of a relevance ordering over possibilities and Lewis's (1996) notion of the elimination of possibilities. Finally, L-semantics follows Lewis's (1996) own $\mathrm{RA}_{\exists\forall}$ theory (for a fixed context).

Definition 3.3 (Truth in an RA Model)**.** Given a relevant alternatives model $\mathcal{M} = \langle W, \sim, \preceq, V\rangle$, a world $w \in W$, and a formula φ in the epistemic language, we define $\mathcal{M}, w \vDash_x \varphi$ (φ is true at w in $\mathcal{M}$ according to X-semantics) as follows:

$$\begin{array}{lll} \mathcal{M}, w \vDash_x p & \text{iff} & w \in V(p)\,; \\ \mathcal{M}, w \vDash_x \neg\varphi & \text{iff} & \mathcal{M}, w \nvDash_x \varphi\,; \\ \mathcal{M}, w \vDash_x \varphi \wedge \psi & \text{iff} & \mathcal{M}, w \vDash_x \varphi \text{ and } \mathcal{M}, w \vDash_x \psi\,. \end{array}$$

For the knowledge operator, the C-semantics clause is:

$$\mathcal{M}, w \vDash_c K\varphi \text{ iff } \forall v \in W : \text{if } w \sim v \text{ then } \mathcal{M}, v \vDash_c \varphi\,,$$

[10] Note that the relevance orderings associated with different worlds may be different. Suppose that w is the actual world, and v is an uneliminated possibility at w with a different relevance ordering, i.e., $\preceq_v \neq \preceq_w$. Our interpretation in this case is that in the actual world, the agent cannot distinguish what is relevant and what is not. As Lewis puts it, "the subject himself may not be able to tell what is properly ignored" (p. 554). This may be because relevance is determined by the conversational context of the ascribers of knowledge, as in Lewis 1996, or by objective features of the agent's situation, as in Dretske 1981.

which states that φ is known at w iff φ is true in all possibilities uneliminated at w. We will write this clause in another, equivalent way, for comparison with the D- and L-semantics clauses. Recall that we interpret $w \prec v$ to mean that v is *ruled out* at w. Then where $[\![\varphi]\!]^{\mathcal{M}}_x = \{v \in W \mid \mathcal{M}, v \vDash_x \varphi\}$ is the set of worlds where φ is true in $\mathcal{M}$ according to X-semantics, the clauses are:

$$\begin{array}{lll} \mathcal{M}, w \vDash_c K\varphi & \text{iff} & \forall v \in [\![\neg\varphi]\!]^{\mathcal{M}}_c : w \prec v\,; \\ \mathcal{M}, w \vDash_d K\varphi & \text{iff} & \forall v \in \mathrm{Max}_{\preceq_w}\left([\![\neg\varphi]\!]^{\mathcal{M}}_d\right) : w \prec v\,; \\ \mathcal{M}, w \vDash_l K\varphi & \text{iff} & \forall v \in \mathrm{Max}_{\preceq_w}(W) \cap [\![\neg\varphi]\!]^{\mathcal{M}}_l : w \prec v\,. \end{array} \quad \begin{array}{l} (\mathrm{RA}_{\forall\exists}) \\ (\mathrm{RA}_{\exists\forall}) \end{array}$$

According to C-semantics, for an agent to know φ at w, *all* $\neg\varphi$-possibilities must be ruled out at w. According to D-semantics, for any φ there is a set of relevant counter-possibilities, namely the set of *most relevant*$_w$ *$\neg\varphi$-possibilities*, $\mathrm{Max}_{\preceq_w}\left([\![\neg\varphi]\!]^{\mathcal{M}}_d\right)$, that an agent must rule out in order to know φ. Finally, according to L-semantics, there is a set of relevant possibilities, $\mathrm{Max}_{\preceq_w}(W)$, such that for any φ, to know φ an agent must rule out the $\neg\varphi$-possibilities *in that set*. Recall the distinction between $\mathrm{RA}_{\forall\exists}$ and $\mathrm{RA}_{\exists\forall}$ introduced above.

Having defined truth according to X-semantics (C/D/L-semantics) we say that φ is *X-valid* iff for all models $\mathcal{M} = \langle W, \sim, \preceq, V\rangle$ and all $w \in W$, $\mathcal{M}, w \vDash_x \varphi$.

To develop a sense for the semantics and how they differ, consider the model in Figure 1 below, drawn for student A in Example 1. An arrow between two worlds indicates that they are indistinguishable for the student according to the relation $\sim$. (Each world is indistinguishable from itself, but we omit reflexive loops.) The ordering of the worlds by their relevance at w_1, which we take to be the actual world, is indicated between worlds.[11] In w_1, the patient has the common condition c, represented by the atomic symbol c. Possibility w_2, in which the patient has the other common condition c' instead of c, is just as relevant$_{w_1}$ (i.e., $w_1 \simeq_{w_1} w_2$). Since the model is for student A, who ran the lab tests to rule out c', she has ruled out w_2 at w_1 (i.e., $w_1 \prec w_2$).[12]

[11] We ignore the relevance orderings associated with other worlds. We also ignore which possibilities are ruled out at worlds other than w_1, since we are not concerned with student A's higher-order knowledge at w_1. If we were, then we would include other worlds in the model. For example, at w_2, all $\neg x$-possibilities are eliminated. But if the patient's condition were c', as in w_2, then student A would still not have ruled out the possibility that the patient has disease x. So we should add a world w'_3, uneliminated at w_2, where x is true.

[12] To be more explicit, we could add new atomic sentences t_c and $t_{c'}$ standing for "the test result indicates c" and "the test result indicates c'," respectively. We would then make t_c true and $t_{c'}$ false at w_1 and w_3, while making $t_{c'}$ true and t_c false at w_2 (and at w'_3, as described in note 11). This would reflect explicitly why w_2 is ruled out at w_1.

A more remote possibility is w_3 (i.e., $w_3 \prec_{w_1} w_2$), in which the patient has the extremely rare disease x. Student A has not run any tests to rule out x, so she has not ruled out w_3 at w_1 (i.e., $w_1 \sim w_3$). Finally, the most remote possibility of all is w_4, in which the patient has both c and x. We assume that student A has learned that x confers immunity to c, so she has ruled out w_4 at w_1.

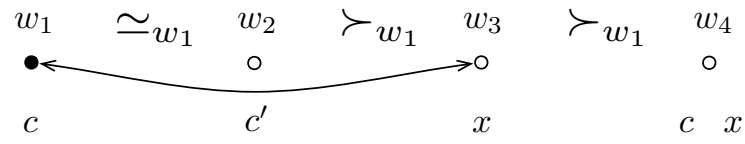


Figure 1: A relevant alternatives model for student A in Example 1.

Now consider C-semantics. In our discussion of Example 1, we held that student A knows that the patient's condition is c, despite the fact that she did not rule out the remote possibility in which the patient has disease x. C-semantics issues the opposite verdict. According to C-semantics, Kc is true at w_1 iff *all* $\neg c$-worlds, regardless of their relevance, are ruled out at w_1. Then since w_3 is not ruled out at w_1, Kc is false at w_1. However, the student has some knowledge at w_1. For example, $K(\neg x \rightarrow c)$ is true at w_1 according to C-semantics.

Next, consider D-semantics. First observe that D-semantics issues our original verdict that student A knows the patient's condition is c. That is, Kc is true at w_1 according to D-semantics, since the most relevant$_{w_1}$ $\neg c$-world, w_2, is ruled out at w_1. $K(c \rightarrow \neg x)$ is also true at w_1, since the most relevant$_{w_1}$ $\neg(c \rightarrow \neg x)$-world, w_4, is ruled out at w_1. Not only that, but $K(c \leftrightarrow \neg x)$ is also true at w_1, since the most relevant$_{w_1}$ $\neg(c \leftrightarrow \neg x)$-world, w_2, is ruled out at w_1. However, the most relevant$_{w_1}$ x-world, w_3, is *not* ruled out at w_1, so $K\neg x$ is false at w_1. Hence student A does not know that the patient does not have disease x.

We have just proven the first part of the following fact, which matches Dretske's (1970) view. The second part matches Lewis's (1996, p. 563, n. 21).

Fact 1. Neither $K\varphi \wedge K(\varphi \rightarrow \psi) \rightarrow K\psi$ nor $K\varphi \wedge K(\varphi \leftrightarrow \psi) \rightarrow K\psi$ are D-valid; but both are C/L-valid.

Proof. For the first part, we have given a countermodel in Figure 1 for both $Kc \wedge K(c \rightarrow \neg x) \rightarrow K\neg x$ and $Kc \wedge K(c \leftrightarrow \neg x) \rightarrow K\neg x$ in D-semantics. The second part follows the the standard proof of the principles in normal modal logic. □

Finally, consider the model in Figure 1 from the perspective of L-semantics. What is noteworthy in this case is that according to L-semantics, student A *does* know that the patient does not have disease x. $K\neg x$ is true at w_1, because $\neg x$ is true in all of the most relevant$_{w_1}$ worlds, namely in w_1 and w_2.

Remark 3.3. The general point is that with L-semantics, an agent can know φ at w even if the agent has not ruled out *any* of the $\neg\varphi$-possibilities. For it may be that none of the $\neg\varphi$-possibilities are relevant at w, i.e., not in $\mathrm{Max}_{\preceq_w}(W)$. This is the position of Stine (1976, p. 257) and Rysiew (2006, p. 265), who hold that one can know that a skeptical possibility does not obtain, even though one cannot rule it out, because the skeptical possibility is not relevant in the context. By contrast, with D-semantics, as long as there is some $\neg\varphi$-possibility, there is some maximally relevant $\neg\varphi$-possibility, which one must rule out to know φ.

Having developed a sense of the semantics, it is straightforward to check that they guarantee what is required for us to call K a knowledge operator, namely that knowledge implies truth. For D- and L-semantics, Fact 2 reflects Lewis's (1996, p. 554) observation that the veridicality of knowledge follows from his Rule of Actuality, given that every world is indistinguishable from itself. Formally, it follows from the fact that w is maximal in $\preceq_w$ and $w \sim w$.

Fact 2 (Veridicality)**.** $K\varphi \to \varphi$ is C/D/L-valid.

In the terminology of Dretske (1970), Fact 1 above shows that the knowledge operator is not a *fully penetrating* operator, since it does not penetrate to all of the logical consequence of what is known. Yet Dretske claims that the knowledge operator is *semi-penetrating*, since it does penetrate to some logical consequences: "it seems to me fairly obvious that if someone knows that P and Q, he thereby knows that Q" and "If he knows that P is the case, he knows that P or Q is the case " (p. 1009). This is supposed to be the "trivial side" of Dretske's thesis (ibid.). However, if we understand the RA theory according to D-semantics, then even these closure principles fail.

Fact 3. $K(\varphi \wedge \psi) \to K\psi$ and $K\varphi \to K(\varphi \vee \psi)$ are not D-valid.

Proof. Figure 1 shows the non-validity of both formulas. For the first, $K(c \wedge \neg x)$ is true at w_1 according to D-semantics, since the most relevant$_{w_1}$ $\neg(c \wedge \neg x)$-world, w_2, is ruled out at w_1. However, we have already seen that $K\neg x$ is false at w_1 according to D-semantics. For the second formula, we have also already seen that Kc is true at w_1 according to D-semantics, yet the most relevant$_{w_1}$ $\neg(c \vee \neg x)$-world, w_3, is uneliminated at w_1, so $K(c \vee \neg x)$ is false at w_1. □

Fact 3 is only the tip of the iceberg, the full extent of which will be revealed in §3.3. Yet it already points to a dilemma. On the one hand, if we understand the RA theory according to D-semantics, then the knowledge operator is not even semi-penetrating, contrary to the "trivial side" of Dretske's thesis. On the other hand, if we understand the theory according to L-semantics, then the knowledge operator is fully-penetrating, contrary to the non-trivial side of Dretske's thesis. It is difficult to escape this dilemma while retaining something like Heller's (1999) picture "of worlds arranged around the actual world in order of similarity, with those that are too far away from the actual world being irrelevant" (p. 199). However, Dretske's (1981) discussion of relevancy sets leaves open whether the RA theory should be developed with this world-ordering picture. In a sequel to this paper I will propose a different way of developing the theory, consistent with Dretske's (1981) discussion, such that the knowledge operator is semi-penetrating, avoiding the dilemma above.

3.2 Counterfactuals and Beliefs

In the previous section, we assumed Lewis's notion of what it is to eliminate a possibility. In this section, we turn to Heller's (1999) notion: "S's ability to rule out not-p be understood thus: S does not believe p in any of the relevant not-p worlds" (p. 198). In contrast to Lewis's notion of elimination, Heller's notion of ruling out applies to alternatives, understood as propositions. In §3.4, we will consider the associated notion of ruling out a possibility.

To capture Heller's notion of ruling out, we define a new class of models, replacing the indistinguishability relation $\sim$ by a *doxastic accessibility* relation $\mathcal{B}$, with which we will represent *belief*. We also relabel the relevance orderings $\preceq_w$ as $\leqslant_w$, where the latter may be interpreted either as relevance orderings or as *similarity* orderings. In the second case, $u \leqslant_w v$ indicates that world v is at least as similar to world w as world u is, in the sense familiar from Lewis 1973.

Definition 3.4 (CB Models)**.** A *counterfactual belief model* is a tuple $\mathcal{M} = \langle W, \mathcal{B}, \leqslant, V\rangle$ where W and V are as in Definition 3.2; $\mathcal{B}$ is a binary relation on W; and $\leqslant$ is a set containing, for each $w \in W$, a similarity relation $\leqslant_w$ that is a total and converse well-founded preorder on W in which w is maximal.

We take $w\mathcal{B}v$ to mean that possibility v is compatible with everything the agent believes in w, and we write $\mathcal{B}(w) = \{v \in W \mid w\mathcal{B}v\}$ for the set of all such *doxastically accessible* worlds from w (see Lewis 1986, §1.4). In Definition 3.5 below, we adopt the standard picture according to which an agent believes φ

at w just in case φ is true throughout her set of doxastically accessible worlds.[13] To express this in our formal language, we add a belief operator B alongside the knowledge operator K, reading $B\varphi$ as "the agent believes that φ."

With the interpretation of $\leqslant_w$ as a similarity ordering, we can capture the following well-known counterfactual conditions on an agent's belief that φ: if φ were false, the agent would not believe φ (*sensitivity*); if φ were true, the agent would believe φ (*adherence*); the agent would believe φ only if φ were true (*safety*). Nozick (1981) argued that sensitivity and adherence are necessary and sufficient for one's belief that φ to constitute knowledge, while Sosa (1999) argued that safety is necessary. (In the full version of this paper, I also consider the revised sensitivity and safety conditions that take into account *methods* of coming to believe and *bases* of belief.) Following Nozick and Sosa, I interpret sensitivity as the counterfactual $\neg\varphi \mathrel{\Box\!\!\rightarrow} \neg B\varphi$, adherence as $\varphi \mathrel{\Box\!\!\rightarrow} B\varphi$, and safety as $B\varphi \mathrel{\Box\!\!\rightarrow} \varphi$. I will understand the truth of counterfactuals following Lewis (1973), such that $\varphi \mathrel{\Box\!\!\rightarrow} \psi$ is true at a world w iff the closest φ-worlds to w according to $\leqslant_w$ (hereafter "closest$_w$") are ψ-worlds.[14]

We now define three semantics for the K operator: H-semantics for **H**eller, N-semantics for **N**ozick, and S-semantics for **S**osa. In doing so, we assume that each theory proposes necessary and sufficient conditions for knowledge. This is true of Nozick's (1981) theory, as it was of Lewis's (1996). However, Sosa (1999) and Heller (1999) propose only necessary conditions. Given this, one may choose to read $K\varphi$ as "the agent *safely believes* φ/has *ruled out the relevant alternatives to* φ" for S/H-semantics. Our results for S/H-semantics can then be viewed as results about the logic of safe belief/the logic of relevant alternatives. (In the full version of this paper, I argue that closure failures for the necessary conditions are very likely to lead to closure failures for knowledge itself, so our results apply to the theories of knowledge of Sosa and Heller after all.)

Definition 3.5 (Truth in a CB Model)**.** Given a counterfactual belief model $\mathcal{M} = \langle W, \mathcal{B}, \leqslant, V\rangle$, a world $w \in W$, and a formula φ in the epistemic-doxastic language, we define $\mathcal{M}, w \vDash_x \varphi$ (φ is true at w in $\mathcal{M}$ according to X-semantics)

[13] This modal model of belief is not essential for any of our results, since they do not deal with higher-order beliefs. We could just as well associate with each world w a consistent, *logically-closed* set of propositional formulas Σ_w such that one believes φ at w just in case $\varphi \in \Sigma_w$. What matters is that our model guarantees the condition from §2 of *full doxastic closure* for the agent (at each world), which the modal model does. See Remark A.1 in the Appendix.

[14] Since the adherence and safety counterfactuals have true antecedents whenever one believes φ, they are trivial in a "centered" model in which each world w is *strictly* maximal in $\leqslant_w$ (see Lewis 1973, §1.7). We will continue to use the weakly centered model (recall note 8) in which it is not redundant, but our results hold for the centered model as well.

as follows (with propositional cases as in Definition 3.3):

$$\mathcal{M}, w \vDash_x B\varphi \quad \text{iff} \quad \forall v \in W: \text{ if } w\mathcal{B}v \text{ then } \mathcal{M}, v \vDash_x \varphi;$$

$$\mathcal{M}, w \vDash_h K\varphi \quad \text{iff} \quad \mathcal{M}, w \vDash_h B\varphi \text{ and } \forall v \in \mathrm{Max}_{\leqslant_w}\left([\![\neg\varphi]\!]_h^{\mathcal{M}}\right): \mathcal{M}, v \nvDash_h B\varphi;$$

$$\begin{aligned}\mathcal{M}, w \vDash_n K\varphi \quad \text{iff} \quad & \mathcal{M}, w \vDash_n B\varphi \text{ and} \\ & \text{(sensitivity) } \forall v \in \mathrm{Max}_{\leqslant_w}\left([\![\neg\varphi]\!]_n^{\mathcal{M}}\right): \mathcal{M}, v \nvDash_n B\varphi, \\ & \text{(adherence) } \forall v \in \mathrm{Max}_{\leqslant_w}\left([\![\varphi]\!]_n^{\mathcal{M}}\right): \mathcal{M}, v \vDash_n B\varphi;\end{aligned}$$

$$\begin{aligned}\mathcal{M}, w \vDash_s K\varphi \quad \text{iff} \quad & \mathcal{M}, w \vDash_s B\varphi \text{ and} \\ & \text{(safety) } \forall v \in \mathrm{Max}_{\leqslant_w}\left([\![B\varphi]\!]_s^{\mathcal{M}}\right): \mathcal{M}, v \vDash_s \varphi.\end{aligned}$$

Note that Heller's (1999) condition for ruling out relevant alternatives is structurally the same as Nozick's (1981) sensitivity condition on belief, though they may interpret the relation $\leqslant_w$ differently (see Heller 1989).

In Figure 2 we display a CB model for Example 1.[15] The arrows now represent the belief relation $\mathcal{B}$. (The dotted arrow only applies for student B, so ignore it for now.) The arrow from w_1 to itself indicates that at w_1, student A believes that the actual situation is w_1. Hence $B\neg x$ is true at w_1. The arrow from w_3 to w_1 indicates that if student A were in w_3, then she would believe that w_1 is the actual situation. (This is because she did not run any of the tests necessary to distinguish c from x.) But then since w_3 is the closest$_{w_1}$ x-world, it follows that if the patient's condition were x, student A would still believe it was c and not x, since c and $\neg x$ are true at w_1. Hence $K\neg x$ is false at w_1 in H/N-semantics, because the sensitivity condition is violated.

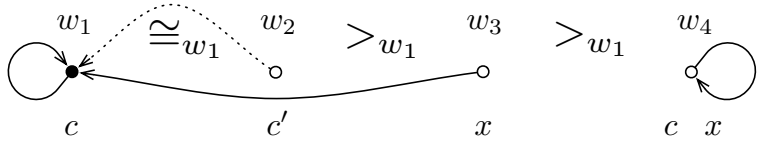


Figure 2: A counterfactual belief model for Example 1.

If we draw the model for student B, including the dotted arrow as well as the solid one, we see that he is in a similar position with respect to the other common condition c'; if the patient's condition were c', he would still believe it

[15]The symbol $\cong_w$ is to $\leqslant_w$ as $\simeq_w$ is to $\preceq_w$; so $v \cong_w u$ means $v \leqslant_w u$ and $u \leqslant_w v$.

was c. (This is because he made the diagnosis of c after only a physical exam, and c and c' have the same visible symptoms.) Kc is false at w_1 in H/N-semantics with the dotted arrow added for student B, but true at w_1 in both semantics if we replace the dotted arrow by an arrow from w_2 to itself for student A.

When we consider S-semantics, we get a different verdict on whether student A knows that the patient does not have disease x. $K\neg x$ is *true* at w_1 in S-semantics, because at the closest$_{w_1}$ worlds, namely w_1 and w_2, student A believes $\neg x$, and $\neg x$ is true at both worlds. Hence student A's belief that $\neg x$ at w_1 is safe and counts as knowledge. Similarly, student A's belief that c at w_1 is safe. Yet if we add the dotted arrow for student B, one can check that student B's belief that c at w_1 is not safe, so Kc is false at w_1 according to S-semantics.

The fact that $K\neg x$ is true at w_1 in S-semantics reflects the idea that the safety theory leads to a neo-Moorean response to skepticism (Sosa 1999), according to which agents can know that skeptical possibilities do not obtain. In general, a point parallel to Remark 3.3 about the RA theory holds for safety: if there are no $\neg\varphi$-worlds among the closest worlds, then one's belief in φ is safe.

Remark 3.4. Alspector-Kelly (2010) remarks that in the definition of safe belief that φ, the set of close worlds does not depend on φ. There is some set of close worlds S, and for any φ, one's belief in φ is safe just in case every $B\varphi$-world in S is a φ-world. This is a $\exists\forall$ definition of safety, as opposed to the $\forall\exists$ definition we have given, which requires that the closest $B\varphi$-worlds satisfy φ. However, this difference is merely apparent. For if $B\varphi$ is true at w, then given that w is maximal in $\leqslant_w$, we have $\mathrm{Max}_{\leqslant_w}\left([\![B\varphi]\!]^{\mathcal{M}}_s\right) = \mathrm{Max}_{\leqslant_w}(W) \cap [\![B\varphi]\!]^{\mathcal{M}}_s$. It follows that we can replace the safety clause in Definition 3.5 by $\forall v \in \mathrm{Max}_{\leqslant_w}(W) : \mathcal{M}, v \vDash_s B\varphi \rightarrow \varphi$, and the resulting truth condition for $K\varphi$ is equivalent. Hence we can consider safety an $\exists\forall$ condition, an important point to which we will return in §3.4.

The proof of the following fact is similar to that of Fact 2.

Fact 4 (Veridicality)**.** $K\varphi \rightarrow \varphi$ is H/N/S-valid.

Next we state a lemma that relates the CB framework of this section to the RA framework of the last. It shows that any counterexample to a closure principle in an RA model under D-semantics can be transferred to a counterexample to the same closure principle in a CB model under H/N-semantics.

Lemma 1.

1. For any RA model $\mathcal{M} = \langle W, \sim, \preceq, V\rangle$ and world $w \in W$, there is a CB model $\mathcal{N} = \langle W, \mathcal{B}, \leqslant, V\rangle$ such that for all propositional formulas φ:

$$\mathcal{M}, w \vDash_d K\varphi \text{ iff } \mathcal{N}, w \vDash_h K\varphi.$$

2. For any CB model $\mathcal{N} = \langle W, \mathcal{B}, \leqslant, V\rangle$ and world $w \in W$, there is a CB model $\mathcal{N}' = \langle W, \mathcal{B}, \leqslant', V\rangle$ such that for all propositional formulas φ:

$$\mathcal{N}, w \vDash_h K\varphi \text{ iff } \mathcal{N}', w \vDash_n K\varphi.$$

We give the proof in the Appendix. It shows how we can relate the notions of ruling out possibilities in RA and CB models, an issue taken up again in §3.4.

It is immediate from Lemma 1 that for any RA model $\mathcal{M}$, if $K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi$ is false at a world w in $\mathcal{M}$, then there is a CB model $\mathcal{N}$ such that $K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi$ is also false at w in $\mathcal{N}$, and similarly for the H- to N-semantics case. Hence the failure of closure under known implication and bi-implication in D-semantics (Fact 1) transfers to H- and N-semantics. The proof of Lemma 1 provides a recipe for building the CB model that falsifies these principles, given the RA model in Figure 1 that falsifies them. In fact, the model in Figure 2 (replacing the dotted arrow by a reflexive loop) is the result.

By Lemma 1 and Fact 3, closure under disjunction introduction and under conjunction elimination also fail in H- and N-semantics. Although Nozick did not explicitly acknowledge that closure under disjunction introduction fails on his theory, that it does fail on his theory can easily be inferred from what he says.[16] On the other hand, Nozick acknowledged that closure under conjunction elimination fails on his theory. In fact, his explanation of why it fails is essentially the same as our proof of the first part of Fact 3.[17]

3.3 The Closure Theorem and Its Consequences

In this section, we state our main result and discuss some of its consequences. Despite the differences between the RA, tracking, and safety theories as formalized, Theorem 1 below provides a unifying perspective: first, the valid epistemic closure principles (recall Definition 3.1) are exactly the same for these different theories of knowledge; second, there is a complete characterization of these closure principles purely in terms of propositional logic.[18]

[16] While Nozick (1981) admits that such a closure failure "surely carries things too far" (p. 230, p. 692, n. 64), he also says that one can know p and yet fail to know $\neg(\neg p \wedge \mathrm{SK})$ (p. 228). But the latter is equivalent to $p \vee \neg\mathrm{SK}$, and Nozick accepts closure under logical equivalence (p. 229).

[17] See the paragraph beginning "S's belief that p&q..." in Nozick 1981, p. 228. Nozick's reasoning reflects the continuity between the nonformal discussions and our formalization, a good sign that the formalization is faithful. As another example, see Vogel 2007, p. 76 for reasoning that is very similar to the proof of Fact 3 for the disjunction case.

[18] Theorem 1 gives a partial answer to the open question, posed by van Benthem (2010, p. 153), of what is the epistemic logic of Nozick's notion of knowledge.

The proof of Theorem 1, given in the Appendix, uses the same sort of reasoning about "closest worlds" that is standard fair in epistemology, only made more mathematical with the help of our formalism. The role of Lemma 1 in the proof is to show that item 2 of the theorem implies item 1, so it only remains to show that 1 implies 3 and that 3 implies both 1 and 2.

Theorem 1 (Closure Theorem)**.** Let $\varphi_1, \ldots, \varphi_n$ and ψ be propositional formulas. The following are equivalent:

1. $K\varphi_1 \wedge \cdots \wedge K\varphi_n \to K\psi$ is D-valid over relevant alternatives models.
2. $K\varphi_1 \wedge \cdots \wedge K\varphi_n \to K\psi$ is H/N/S-valid over counterfactual belief models.
3. One of the following is the case:
   (a) ψ is a tautology;
   (b) $\varphi_1 \wedge \cdots \wedge \varphi_n$ is a contradiction;
   (c) There are $\varphi_1, \ldots, \varphi_m$ among $\varphi_1, \ldots, \varphi_n$ such that $\varphi_1 \wedge \cdots \wedge \varphi_m \leftrightarrow \psi$ is a tautology.

Consider how the claim of Theorem 1 applies to our earlier examples: $Kp \wedge K(p \to q) \to Kq$ is not D/H/N/S-valid, because $p \wedge (p \to q) \leftrightarrow q$ is not a tautology, and neither is $p \leftrightarrow q$ or $(p \to q) \leftrightarrow q$; similarly, $K(p \wedge q) \to Kp$ is not valid, because $p \wedge q \leftrightarrow p$ is not a tautology; and $Kp \to K(p \vee q)$ is not valid, because $p \leftrightarrow p \vee q$ is not a tautology. On the other hand, we now see that $Kp \wedge Kq \to K(p \wedge q)$ *is* D/H/N/S-valid, because $p \wedge q \leftrightarrow p \wedge q$ is a tautology.

Theorem 1 precisely charts the extent of closure failure—the Problem of Containment—for the basic subjunctivist approach to knowledge, where I now take the label 'subjunctivist' to apply not only to N- and S-semantics, but also to D- and H-semantics, given their structural similarities. Note that since we are considering the knowledge of an ideally astute logician, in the sense explained in §2, these closure failures are not due to the agent's failure to believe the logical consequences of what she know. Rather, they are due to the agent's satisfying the conditions for knowledge (ruling out the relevant alternatives, tracking the truth, etc.) with respect to some propositions and yet not with respect to all logical consequences of the set of those propositions, even though the agent believes all of the consequences. (In the full version of this paper, I argue that the closure failures do not go away even if we assume that the agent has come to believe these consequences *by deduction* and even if we build the method or basis of an agent's beliefs into the conditions of knowledge.)

For some instances of serious closure failure, it follows from Theorem 1 that none of the following are D/H/N/S-valid:

(i) $K(\varphi \wedge \psi) \rightarrow K\varphi$; (iii) $K(\varphi \wedge \psi) \rightarrow K(\varphi \vee \psi)$;
(ii) $K\varphi \rightarrow K(\varphi \vee \psi)$; (iv) $K\varphi \wedge K\psi \rightarrow K(\varphi \vee \psi)$.

We have already discussed the failures of principles (i) and (ii) in Fact 3. Now we see that even (iii) and (iv) fail for subjunctivists. But (iii) is intuitively an even weaker closure principle than (i) or (ii). For the antecedent of (iii) is the same as that of (i), but the consequent of (iii) is intuitively weaker than that of (i); and the consequent of (iii) is the same as that of (ii), but the antecedent of (iii) is intuitively stronger than that of (ii). Principle (iv) is weaker still, if we take the perspective of the subjunctivist semantics, according to which the antecedent of (iv) implies that of (iii) (as noted above), but not vice versa.

Recall from §2 that the case against closure under known implication (**K**) had two components: examples of situations in which the principle appears (to some) to fail and a theory that purports to explain the failures. For (iii) and (iv), we have only a theoretical explanation. Even in the case of (ii), there are examples of "junk disjunctive knowledge" (Hawthorne 2004, 71ff) in which it may appear that (ii) fails. However, such examples do not apply to (iii) or (iv). Without some convincing examples in which (iii) and (iv) fail, their failure according to a theory should count as strong evidence against the theory—even for those who are otherwise sympathetic to the denial of **K**.

While failures of closure spread too far given subjunctivism, they also do not spread *far enough* for those who wish to avoid skepticism by denying closure. Let p be an ordinary proposition and SK a skeptical hypothesis incompatible with p. Say that one is in the *skeptical predicament* (SP) when Kp and $K(p \rightarrow \neg\mathrm{SK})$ are true but $K\neg\mathrm{SK}$ is false, which violates **K**. According to subjunctivists, SP is possible. Indeed, this is the crux of Dretske and Nozick's defense against skepticism. However, by Theorem 1, while **K** is not D/H/N/S-valid, both

(v) $(K(\varphi \vee \psi) \wedge K(\varphi \rightarrow \psi)) \rightarrow K\psi$ and (vi) $(K\varphi \wedge K(\varphi \rightarrow \psi)) \rightarrow K(\varphi \wedge \psi)$

are D/H/N/S-valid. Yet the validity of (v) and (vi) seems to undermine any anti-skeptical motivation that one might have for denying **K**. Given the validity of (v), the subjunctivist must insist that whenever an agent is in SP, although Kp is true, $K(p \vee \neg\mathrm{SK})$ is *false*; for otherwise $K\neg\mathrm{SK}$ would be true by (v), contrary to the description of SP. Therefore, in order to defend against skepticism by claiming that **K** fails, subjunctivists must also claim that closure under disjunction introduction (in this instance, $Kp \rightarrow K(p \vee \neg\mathrm{SK})$) fails, in conflict with Nozick

(1981, p. 692, n. 64) and Dretske's (1970, p. 1009) endorsement of that closure principle. Not only that, but given the validity of (vi), $K(p \wedge \neg \mathrm{SK})$ is always *true* when an agent is in SP, even though $K\neg \mathrm{SK}$ is false. But just as the skeptic argues by *modus tollens* from $\neg K\neg \mathrm{SK}$ and **K** to the conclusion of $\neg Kp$, should the skeptic not just as well argue from $\neg K(p \wedge \neg \mathrm{SK})$ and (vi) to the same conclusion?[19]

More could be said about specific closure principles that fail or hold on the subjunctivist semantics, but we leave this to the reader.[20] We also leave it to the reader to check that in C/L-semantics, all closure principles are valid.[21]

3.4 Theory Parameters and Closure

Analysis of Theorem 1 shows that two parameters of a modal theory of knowledge affect whether closure holds. We have already identified one: the choice of the relevancy set. A theory has an $\exists\forall$ setting of this parameter if there is some set of relevant worlds, which does not depend on the φ in question, such that to know φ one must meet some condition (e.g., ruling out the $\neg\varphi$-worlds or believing φ only when φ is true) with respect to those worlds. L- and S-semantics have an $\exists\forall$ setting in this sense (recall Remark 3.4). By contrast, a theory has a $\forall\exists$ setting of this parameter if for each φ, there is a set of relevant-to-φ worlds such that to know φ one must meet some condition with respect to those worlds. D- and H/N-semantics have a $\forall\exists$ setting in this sense.[22]

It is noteworthy that both L- and S-semantics have an $\exists\forall$ choice of relevancy sets, and yet closure holds in L-semantics but fails in S-semantics. The explanation for this difference involves the second theory parameter: the notion of ruling out. According to the (Lewisian) notion of ruling out used in L- and D-semantics, a world v is either ruled out at w or not. For either v is indistinguishable from w or it is not. By contrast, according to the notion of ruling out implicit in S-, H-, and N-semantics, we cannot say independently of the proposition in question whether a world v is ruled out at w or not.

[19] Nozick (1981, p. 229) does not think so, since he holds that the agent does know $p \wedge \neg \mathrm{SK}$, but not $\neg \mathrm{SK}$. He seems not to have realized the point about disjunction and (v).

[20] Theorem 1 shows that the valid closure principles for D-, H-, N-, and S-semantics are exactly the same. Differences only appear when we allow disjunctions in the consequents of conditionals. For example, one can check that $K(\varphi \wedge \psi) \rightarrow (K\varphi \vee K\psi)$ is D-valid, but not H/N-valid.

[21] By basic modal logic, all closure principles are C-valid. But then they are also L-valid, for if there were a countermodel $\mathcal{M}, w$ in L-semantics to a closure principle, we could obtain a countermodel $\mathcal{M}', w$ to the same closure principle in C-semantics by eliminating from $\mathcal{M}$ any worlds that are not maximally relevant at w (and restricting the relations and valuation accordingly). Cf. Lemma 1.

[22] The sensitivity condition has the $\forall\exists$ character. By contrast, adherence has an $\exists\forall$ character, and similar remarks apply to adherence as we made for safety in Remark 3.4.

Consider the CB model in Figure 3, where the arrows represent the belief relation $\mathcal{B}$ as before. In this model, the closest$_{w_1}$ worlds are w_1 and w_2. We may say that w_2 is ruled out at w_1 *with respect to* $p \wedge q$, in the sense that while $p \wedge q$ is false at w_2, the agent does not believe $p \wedge q$ at w_2 (but rather $\neg p \wedge q$). However, w_2 is not ruled out at w_1 *with respect to* q, for q is false at w_2 and yet the agent believes q at w_2. This reflects what we already know from Theorem 1, namely that one's belief that $p \wedge q$ may be safe at w_1 while one's belief that q is not safe at w_1, which is why $K(p \wedge q) \rightarrow Kq$ is not valid in S-semantics.

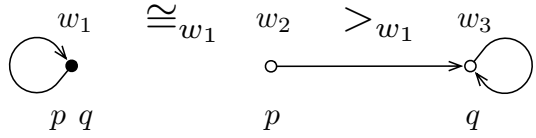


Figure 3: A CB countermodel to $K(p \wedge q) \rightarrow Kq$ in S-semantics.

The distinction between the two notions of ruling out is again $\exists\forall$ vs. $\forall\exists$:

- In L/D-semantics, for a given world v, there exists ($\exists$) a *ruled out status* for v at w, which holds for all ($\forall$) φ.
- In S/H/N-semantics, for a given world v, for every ($\forall$) φ there exists ($\exists$) a ruled out *with respect to* φ status for v at w.

To grasp the distinction, an explicit parallel is helpful: given a $\forall\exists$ setting of the ruling out (resp. relevancy set) parameter, a $\neg\varphi \wedge \neg\psi$-world that is "ruled out" with respect to φ (resp. that must be ruled out in order to know φ) may not be "ruled out" with respect to ψ (resp. may not be such that it must be ruled out in order to know ψ), since the choice of the world's status as ruled out or not (resp. as most relevant or not) depends on the proposition, as indicated by the $\forall$ *propositions* $\exists$ *ruled out status* (resp. $\exists$ *relevancy set*) quantifier dependence. As the example in Figure 3 shows, the $\forall\exists$ notion of ruling out explains why closure fails in S-semantics, despite its $\exists\forall$ setting of the relevancy set parameter.

Table 1 below displays the relationship between the two theory parameters and closure failures. This analysis suggests the following informal conjecture.

Conjecture 3.1. In order for a modal theory of knowledge to support full epistemic closure, an $\exists\forall$ setting of both theory parameters is necessary.

Theory	Formalization	Relevancy Set	Ruling Out	Closure Failures
RA	L-semantics	$\exists\forall$	$\exists\forall$	none
RA	D-semantics	$\forall\exists$	$\exists\forall$	Theorem 1
Safety	S-semantics	$\exists\forall$	$\forall\exists$	Theorem 1
Tracking	H/N-semantics	$\forall\exists$	$\forall\exists$	Theorem 1

Table 1: Parameter Settings and Closure Failures

4 Conclusion

Having discussed our main results in §3.3 - 3.4, we end with a summary of the topics treated in the full version of this preprint, as well as in the papers to follow in this series, and with some brief reflections on methodology.

In the full version of this preprint, I argue that modified versions of the basic subjunctivist theories studied here do not avoid the Problem of Containment with which we began in §1. In particular, DeRose's (1995) modified version of the tracking theory suffers from many of the same serious closure failures as Nozick's original version, and versions of the tracking and safety theories that take into account *methods* or *bases* of belief do not avoid closure failures (cf. Alspector-Kelly 2010 on safety). Moreover, even if the subjunctivist conditions on belief are taken only as necessary conditions for knowledge, closure failures for these necessary conditions are very likely to result in closure failures for knowledge itself (cf. Brueckner 2004 and Murphy 2006). Finally, I show how the structural features of subjunctivist theories that lead to failures of epistemic closure also lead to problems of higher-order knowledge.

In the sequel to this paper, I generalize the formal framework developed here, in order to use it as a guide to a *solution* of the Problem of Containment. In particular, I argue that there is a natural version of the RA theory for which "dangerous" epistemic closure principles that lead to skepticism fail, while innocuous closure principles hold. According to this new theory, the knowledge operator is semi-penetrating, as Dretske (1970) desired, and the dilemma raised at the end of §3.1 is resolved. Finally, in the third installment of this series, I extend the formal framework of the first two papers in order to model *contextualist* versions of the RA (Lewis 1996) and tracking (DeRose 1995) theories. After developing the resources to model the dynamics of *context change*, I raise

doubts about whether these contextualist theories can fulfill their promise to handle skepticism while preserving closure in a significant sense.

The results of this paper already reveal noteworthy features of our epistemic-logical approach. In epistemology, an indispensable method of theory assessment begins by considering the verdicts issued by different theories about which knowledge claims are true in a particular scenario. This is akin to considering the verdicts issued by different semantics about which knowledge formulas are true in a particular model. All of the semantics we studied can issue different verdicts for the same model. Moreover, theorists who favor different semantics may represent a scenario with different models in the first place. However, what we have seen is that by rising to the level of *truth in all models*, of validity, these differences may wash away, revealing unity on a higher level. Theorem 1 provided such a perspective, by showing that four different epistemological theories validate exactly the same epistemic closure principles.

For some philosophers, a source of hesitation about epistemic logic is the level of idealization. In basic systems of epistemic logic, agents know all valid principles of the logic, and they know all the logical consequences of what they know. This is the much-discussed "problem of logical omniscience" (Stalnaker 1991). Yet in our setting, logical omniscience is a feature, not a bug. Although in our formalizations of the RA and subjunctivist theories, agents do not know all the logical consequences of what they know, given failures of epistemic closure, they are still logically omniscient in another sense. For they know all tautologies, and they believe all the logical consequences of what they believe. Not only are these properties desirable but they are arguably necessary if we are to distinguish failures of epistemic closure that are due to bounded deductive power from failures of closure that are due to the nature of knowledge according to the RA and subjunctivist theories.[23] This shows the positive role that idealization can play in epistemology, as it does in science.

Acknowledgements For helpful discussions and comments on this paper, I wish to thank Johan van Benthem, Tomohiro Hoshi, Thomas Icard, Alistair Isaac, Krista Lawlor, Neil van Leeuwen, Helen Longino, and Eric Pacuit.

[23] Williamson (2010) makes a similar point that logical omniscience is a desirable assumption if we are to discern the specific effects of limited powers of discrimination on knowledge.

A Appendix

For the following proofs, we refer to worlds in $\mathrm{Max}_{\preceq_w}([\![\varphi]\!]^{\mathcal{M}})$ as *closest*$_w$ *φ-worlds* and worlds in $\mathrm{Max}_{\preceq_w}(W)$ as *closest*$_w$ *worlds*, and similarly for $\leqslant_w$. If φ is true at a world w in X-semantics, then we say that φ is *X-true* at w.

Lemma 1.

1. For any RA model $\mathcal{M} = \langle W, \sim, \preceq, V\rangle$ and world $w \in W$, there is a CB model $\mathcal{N} = \langle W, \mathcal{B}, \leqslant, V\rangle$ such that for all propositional formulas φ:

$$\mathcal{M}, w \vDash_d K\varphi \text{ iff } \mathcal{N}, w \vDash_h K\varphi.$$

2. For any CB model $\mathcal{N} = \langle W, \mathcal{B}, \leqslant, V\rangle$ and world $w \in W$, there is a CB model $\mathcal{N}' = \langle W, \mathcal{B}, \leqslant', V\rangle$ such that for all propositional formulas φ:

$$\mathcal{N}, w \vDash_h K\varphi \text{ iff } \mathcal{N}', w \vDash_n K\varphi.$$

Proof. For part 1, we construct $\mathcal{N}$ from $\mathcal{M}$ as follows. Let the set of worlds W and valuation V in $\mathcal{N}$ be the same as those in $\mathcal{M}$; let the set of similarity relations $\leqslant$ in $\mathcal{N}$ be the same as the set of relevance relations $\preceq$ in $\mathcal{M}$; finally, construct the doxastic accessibility relation $\mathcal{B}$ in $\mathcal{N}$ from the indistinguishability relation $\sim$ in $\mathcal{M}$, such that if $w \sim v$ in $\mathcal{M}$, then $\mathcal{B}(v) = \{w\}$, and otherwise $\mathcal{B}(v) = \{v\}$.

To show the equivalence in part 1, suppose $K\varphi$ is D-true at $\mathcal{M}, w$. Hence φ is true at $\mathcal{M}, w$ by veridicality (Fact 2). Since φ is propositional and the valuation of $\mathcal{N}$ is the same as that of $\mathcal{M}$, it follows that φ is true at $\mathcal{N}, w$. Then since $w \sim w$, we have $\mathcal{B}(w) = \{w\}$ by construction of $\mathcal{N}$, in which case $B\varphi$ is true at $\mathcal{N}, w$. Having shown that knowledge at $\mathcal{M}, w$ implies belief at $\mathcal{N}, w$, to complete both directions of part 1 it only remains to establish the fact (\$) that $K\varphi$ is D-true at $\mathcal{M}, w$ iff the sensitivity condition for $K\varphi$ is satisfied at $\mathcal{N}, w$.

Given that the set of worlds, valuation, and relevance/similarity relations are the same in $\mathcal{N}$ and $\mathcal{M}$, and that φ is propositional, we have that the closest$_w$ $\neg\varphi$-worlds are the same in $\mathcal{M}$ (according to $\preceq_w$) as in $\mathcal{N}$ (according to $\leqslant_w$). If φ is false at w in $\mathcal{M}$ and hence at w in $\mathcal{N}$, then $K\varphi$ is false at both by veridicality (Facts 2 and 4), so we are done. If φ is true at both, then it suffices to show that for all closest$_w$ $\neg\varphi$-worlds u, we have $w \nsim u$ in $\mathcal{M}$ iff $B\varphi$ is false at $\mathcal{N}, u$, i.e., the closest$_w$ $\neg\varphi$-worlds are ruled out at w in the D-semantics sense iff they are ruled out in the H-semantics sense, from which (\$) above follows. By construction of $\mathcal{N}$, we have $w \nsim u$ iff $\mathcal{B}(u) = \{u\}$. Then if u is a $\neg\varphi$-world, we have $\mathcal{B}(u) = \{u\}$ iff $B\varphi$ is false at $\mathcal{N}, u$. The left-to-right direction is by the truth definition for

belief. For the contrapositive of the right-to-left direction, if $\mathcal{B}(u) \neq \{u\}$, then $\mathcal{B}(u) = \{w\}$ by construction of $\mathcal{N}$, in which case $B\varphi$ is true at $\mathcal{N}, u$ given the assumption that φ is true at $\mathcal{N}, w$. Therefore, for all closest$_w$ $\neg\varphi$-worlds u, we have $w \nsim u$ iff $\mathcal{B}(u) = \{u\}$ iff $B\varphi$ is false at $\mathcal{N}, u$, as desired.

For part 2, we construct $\mathcal{N}'$ from $\mathcal{N}$ by a simple modification. We make w *strictly* maximal in $\leqslant'_w$, but otherwise keep everything the same, i.e., for all $u, v \in W$, if $u \neq w$, then $u <'_w w$; if $u \neq w$ and $v \neq w$, then $u \leqslant'_w v$ iff $u \leqslant_w v$. We keep the similarity relations $\leqslant_v$ for the other worlds $v \neq w$ exactly the same.

We leave the right-to-left direction of part 2 to the reader. For the left-to-right direction, suppose $K\varphi$ is H-true at $\mathcal{N}, w$. Hence φ is true at $\mathcal{N}, w$ by veridicality (Fact 4). Since φ is propositional and the valuation of $\mathcal{N}'$ is the same as that of $\mathcal{N}$, it follows that φ is true at $\mathcal{N}', w$. Then by the definition of $\leqslant'_w$, we have the fact ($\sharp$) that the closest$_w$ $\neg\varphi$-worlds are the same in $\mathcal{N}'$ (according to $\leqslant'_w$) as in $\mathcal{N}$ (according to $\leqslant_w$). Since $K\varphi$ is H-true at $\mathcal{N}, w$, we have that in $\mathcal{N}$, $B\varphi$ is true at w but false at all of the closest$_w$ $\neg\varphi$-worlds; and given ($\sharp$) and the fact that the doxastic relation $\mathcal{B}$ is the same in $\mathcal{N}'$ as in $\mathcal{N}$, the previous fact also holds in $\mathcal{N}'$. Hence the belief and sensitivity conditions for knowledge of φ are satisfied at $\mathcal{N}', w$. Finally, since w is the *strictly* closest$_w$ $B\varphi$-world in $\mathcal{N}'$ and φ is true at $\mathcal{N}', w$, the adherence condition is also satisfied, so $K\varphi$ is N-true at $\mathcal{N}', w$. □

Remark A.1. For the following proof, we use the basic fact (see, e.g., Theorem 3.3(2) of Chellas 1980) that the semantics for belief given in §3.2 guarantees that if $\alpha_1 \wedge \cdots \wedge \alpha_k \rightarrow \beta$ is valid according to H-, N-, or S-semantics, then $B\alpha_1 \wedge \cdots \wedge B\alpha_k \rightarrow B\beta$ is valid according to H-, N-, or S-semantics, respectively, reflecting our assumption of *full doxastic closure* from §2.

Theorem 1 (Closure Theorem). Let $\varphi_1, \ldots, \varphi_n$ and ψ be propositional formulas. The following are equivalent:

1. $K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi$ is D-valid over relevant alternatives models.
2. $K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi$ is H/N/S-valid over counterfactual belief models.
3. One of the following is the case:
   (a) ψ is a tautology;
   (b) $\varphi_1 \wedge \cdots \wedge \varphi_n$ is a contradiction;
   (c) There are $\varphi_1, \ldots, \varphi_m$ among $\varphi_1, \ldots, \varphi_n$ such that $\varphi_1 \wedge \cdots \wedge \varphi_m \leftrightarrow \psi$ is a tautology.

Proof. 3 $\Rightarrow$ (1 & 2). (a) If ψ is a tautology, then in any RA/CB model $\mathcal{M}$, there are no $\neg\psi$-worlds, in which case $K\psi$ is true at any world w in $\mathcal{M}$ under D/H/N/S-semantics. (b) If $\varphi_1 \wedge \cdots \wedge \varphi_n$ is a contradiction, then given veridicality, $K\varphi_1 \wedge \cdots \wedge K\varphi_n$ is unsatisfiable, in which case $K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi$ is valid. (c) It suffices to show that if $\varphi_1 \wedge \cdots \wedge \varphi_m \leftrightarrow \psi$ is a tautology, then $K\varphi_1 \wedge \cdots \wedge K\varphi_m \rightarrow K\psi$ is D-valid over RA models and H/N/S-valid over CB models. (Then it follows by strengthening the antecedent that $K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi$ is also valid.) For the D-, H-, and N-semantics cases, we do so using the following fact:

> (†) If $\varphi_1 \wedge \cdots \wedge \varphi_m \leftrightarrow \psi$ is a tautology, then for any RA/CB model $\mathcal{M}$ and world w in $\mathcal{M}$, it holds that for every closest$_w$ $\neg\psi$-world s in $\mathcal{M}$, there is some $i \leq m$ such that s is a closest$_w$ $\neg\varphi_i$-world in $\mathcal{M}$.

Before proving (†), we show that given (†), (c) implies 1 and 2.

(†) $\Rightarrow$ ((*c*) $\Rightarrow$ 1). Consider an RA model $\mathcal{M}$ and world w in $\mathcal{M}$ such that $K\varphi_i$ is D-true at w for all $i \leq n$. It follows by the D-truth definition that the closest$_w$ $\neg\varphi_i$-worlds are ruled out at w for all $i \leq m$. Then by (†) and (c), the closest$_w$ $\neg\psi$-worlds are also ruled out at w, so $K\psi$ is D-true at w.

(†) $\Rightarrow$ ((*c*) $\Rightarrow$ 2). We use the fact ($\triangle$) that if (c) holds and hence $\psi \rightarrow \varphi_i$ is valid for all $i \leq m$, then $\neg B\varphi_i \rightarrow \neg B\psi$ is also valid for all $i \leq m$ by Remark A.1 above. Consider a CB model $\mathcal{M}$ and world w in $\mathcal{M}$ such that $K\varphi_i$ is H-true at w for all $i \leq n$. It follows by the H-truth definition that for all $i \leq m$, the closest$_w$ $\neg\varphi_i$-worlds satisfy $\neg B\varphi_i$. Then by (†), (c), and ($\triangle$), the closest$_w$ $\neg\psi$-worlds satisfy $\neg B\psi$, so the sensitivity condition holds and $K\psi$ is H-true at w.

To show that $K\psi$ is N-true at w, we must also show that the adherence condition holds: all closest$_w$ ψ-worlds satisfy $B\psi$. Consider such a closest$_w$ ψ-world v. Since $K\psi$ is H-true at w, ψ is true at w by veridicality. Hence v must be one of the closest$_w$ worlds. Then given that $\psi \rightarrow \varphi_i$ is valid for all $i \leq m$ by (c), v is a closest$_w$ φ_i-world for all $i \leq m$. Since we are assuming that for all $i \leq n$, $K\varphi_i$ is N-true at w, by the adherence condition we have that for all $i \leq m$, $B\varphi_i$ is true at the closest$_w$ φ_i-worlds; hence for all $i \leq m$, $B\varphi_i$ is true at v. Finally, by (c) again, $\varphi_i \wedge \cdots \wedge \varphi_m \rightarrow \psi$ is valid, so $B\psi$ is true at v by Remark A.1. Since v was arbitrary, all of the closest$_w$ ψ-worlds satisfy $B\psi$, as desired.

We prove (†) itself by *reductio*. Suppose (i) $\varphi_1 \wedge \cdots \wedge \varphi_m \leftrightarrow \psi$ is a tautology, but (ii) for some RA/CB model $\mathcal{M}$ and world w in $\mathcal{M}$, there is a closest$_w$ $\neg\psi$-world that is not a closest$_w$ $\neg\varphi_i$-world for any $i \leq m$. By the assumption that the ordering $\preceq_w$ (resp. $\leqslant_w$) is converse well-founded, there must be a closest$_w$ $\neg\varphi_j$-world v for some $j \leq m$, such that there is no $\neg\varphi_i$-world for any $i \leq m$ that is closer$_w$ than v. Given (i), v is a $\neg\psi$-world. Then given (ii), there is a closest$_w$

$\neg\psi$-world u that is at least as close$_w$ as v and that is not a closest$_w$ $\neg\varphi_i$-world for any $i \leq m$. However, since u is a $\neg\psi$-world, by (i) it is also a $\neg\varphi_k$-world for some $k \leq m$, in which case by the definition of v it follows that u is a *closest*$_w$ $\neg\varphi_k$-world, a contradiction. This completes the proof of (†).

For S-semantics, suppose (c) holds, and consider a CB model $\mathcal{M}$ and world w in $\mathcal{M}$ such that $K\varphi_i$ is S-true at w for all $i \leq n$. It follows by the S-truth definition that $B\varphi_i$ is true at w for all $i \leq m$. Then given that $\varphi_1 \wedge \cdots \wedge \varphi_m \rightarrow \psi$ is valid, $B\psi$ is true at w by Remark A.1. Hence the closest$_w$ $B\psi$-worlds are among the closest$_w$ worlds. Consider such a closest$_w$ $B\psi$-world v. Given that $\psi \rightarrow \varphi_1 \wedge \cdots \wedge \varphi_m$ is valid, by Remark A.1, v is a $B\varphi_i$-world—indeed, a closest$_w$ $B\varphi_i$-world—for all $i \leq m$. Since we are assuming that $K\varphi_i$ is S-true at w for all $i \leq n$, it follows by the S-truth definition that for all $i \leq m$, φ_i is true at the closest$_w$ $B\varphi_i$-worlds. Hence for all $i \leq m$, φ_i is true at v. Finally, given that $\varphi_1 \wedge \cdots \wedge \varphi_m \rightarrow \psi$ is valid, ψ is true at v. Since v was arbitrary, all closest$_w$ $B\psi$-worlds satisfy ψ, so the safety condition holds and $K\psi$ is S-true at w.

(1 ⇒ 3) & (2 ⇒ 3). For the D-, H-, and N-semantics cases, we prove:

> (‡) If neither (a) nor (b) nor (c) holds, then there is an RA countermodel for $K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi$ in D-semantics.

By Lemma 1, if there is an RA countermodel in D-semantics, then there are also CB countermodels in H- and N-semantics, so for the H and N cases, it suffices to prove (‡). To that end, assume that (a), (b), and (c) do not hold. In the following, we use a compact notation for representing countermodels.[24]

Case 1: $\nvDash \varphi_1 \wedge \cdots \wedge \varphi_n \rightarrow \psi$.[25] Then $\underline{(\varphi_1, \ldots, \varphi_n, \neg\psi)}$ is a countermodel for $K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi$.

Case 2: $\vDash \varphi_1 \wedge \cdots \wedge \varphi_n \rightarrow \psi$. In this case, we prove by induction on n that if (a), (b), and (c) do not hold for $K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi$, and $\vDash \varphi_1 \wedge \cdots \wedge \varphi_n \rightarrow \psi$, then there is a countermodel for $K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi$.

---

[24] We represent worlds by parentheses $(\ldots)$, and we write (α, β, γ) to indicate that propositional formulas α, β, and γ are true at the represented world. Listing worlds in a row, $(\ldots) > (\ldots)$, indicates that the left world is closer than the right world in the relevance ordering *of the leftmost world in the row*, which we take to be the actual world at which the closure formulas are evaluated. Underlining, as in $\underline{(\ldots)}$, indicates that the underlined world is uneliminated at the actual world. When the existence of a world with a particular valuation is given by an assumption (x), we write $(\ldots)^{(x)}$. We leave it to the reader to verify that the models presented are countermodels as claimed.

[25] We use the standard notation $\nvDash \alpha$ to indicate that α is *not valid* and $\vDash \alpha$ to indicate that α is valid. Since the $\varphi_1, \ldots, \varphi_n$ and ψ are propositional, this is just propositional validity.

Base case. Suppose (a), (b), and (c) do not hold for $K\varphi_1 \to K\psi$, and $\vDash \varphi_1 \to \psi$. Since (c) does not hold, $\vDash \varphi_1 \to \psi$ implies $\nvDash \psi \to \varphi_1$ (*i*). Then given that (a) and (b) do not hold, the countermodel for $K\varphi_1 \to K\psi$ is: $\underline{(\varphi_1)} > (\neg\varphi_1, \psi)^{(i)} > \underline{(\neg\psi)}$.

Inductive step. Suppose that (a), (b), and (c) do not hold for $K\varphi_1 \wedge \cdots \wedge K\varphi_{n+1} \to \psi$, and $\vDash \varphi_1 \wedge \cdots \wedge \varphi_{n+1} \to \psi$. By the first assumption, we have the fact (∇) that (a), (b), and (c) do not hold for $K\varphi_1 \wedge \cdots \wedge K\varphi_n \to K\psi$. Since (c) does not hold for $K\varphi_1 \wedge \cdots \wedge K\varphi_{n+1} \to \psi$, it follows from the second assumption that $\nvDash \psi \to \varphi_1 \wedge \cdots \wedge \varphi_{n+1}$, in which case $\nvDash \psi \to \varphi_k$ for some $k \leq n+1$. Without loss of generality, suppose $\nvDash \psi \to \varphi_{n+1}$ (*ii*).

Case 2.1: $\nvDash \varphi_1 \wedge \cdots \wedge \varphi_n \to \psi$ (*iii*). Then the countermodel for $K\varphi_1 \wedge \cdots \wedge K\varphi_{n+1} \to K\psi$ is: $\underline{(\varphi_1, \ldots, \varphi_{n+1})} > (\neg\varphi_{n+1}, \psi)^{(ii)} > \underline{(\varphi_1, \ldots, \varphi_n, \neg\psi)}^{(iii)}$.

Case 2.2: $\vDash \varphi_1 \wedge \cdots \wedge \varphi_n \to \psi$. Together with ($\nabla$), this implies by the inductive hypothesis that there is a countermodel $\mathcal{M}, w$ for $K\varphi_1 \wedge \cdots \wedge K\varphi_n \to K\psi$.

To obtain a countermodel $\mathcal{M}^*, w$ for $K\varphi_1 \wedge \cdots \wedge K\varphi_{n+1} \to K\psi$, we transform $\mathcal{M}$ in three steps. First, we transform $\mathcal{M} = \langle W, \sim, \leq, V\rangle$ into $\mathcal{M}' = \langle W, \sim, \leq', V\rangle$ by making w *strictly* maximal in $\leq'_w$, as in the proof of Lemma 1.2. We claim that $\mathcal{M}', w$ is still a countermodel for $K\varphi_1 \wedge \cdots \wedge K\varphi_n \to K\psi$. Given that $\mathcal{M}, w$ is a countermodel, by veridicality we have that $\varphi_1, \ldots, \varphi_n$ are true at w, in which case so is ψ by the assumption of the case. Hence the $\text{closest}_w\ \neg\varphi_i$-worlds for $i \leq n$ and the $\text{closest}_w\ \neg\psi$-worlds do not change from $\mathcal{M}$ to $\mathcal{M}'$, and neither does the set of worlds that are ruled out at w. It follows that $\mathcal{M}', w$ is a countermodel given that $\mathcal{M}, w$ is. In the second step, we transform $\mathcal{M}'$ to $\mathcal{M}'' = \langle W, \sim, \leq', V''\rangle$ by changing the valuation at w such that $\varphi_1, \ldots, \varphi_{n+1}$ (and hence ψ) are true at w, which is possible given that (b) does not hold. Then it is easy to see that $\mathcal{M}'', w$ is still a countermodel for $K\varphi_1 \wedge \cdots \wedge K\varphi_n \to K\psi$. Finally, in the third step, we transform $\mathcal{M}''$ to $\mathcal{M}^* = \langle W^*, \sim^*, \leq^*, V^*\rangle$ by adding to W a $(\neg\varphi_{n+1}, \psi)^{(ii)}$ world v and setting $w \simeq^*_w v$ but $w \not\sim^* v$;[26] otherwise $\leq^*$, $\sim^*$, and V^* agree with $\leq'$, $\sim'$, and V'', respectively, on W. Then since in $\mathcal{M}''$, all $\text{closest}_w\ \neg\varphi_i$-worlds are ruled out for all $i \leq n$, in $\mathcal{M}^*$ all $\text{closest}_w\ \neg\varphi_i$-worlds are ruled out for all $i \leq n+1$. However, in $\mathcal{M}''$ there is a $\text{closest}_w\ \neg\psi$-world that is not ruled out at w, and this world is also a $\text{closest}_w\ \neg\psi$-world that is not ruled out at w in $\mathcal{M}^*$. Hence $\mathcal{M}^*, w$ is a countermodel for $K\varphi_1 \wedge \cdots \wedge K\varphi_{n+1} \to K\psi$, as desired.

The proof by induction is complete.

For the S-semantics case, assume that (a), (b), and (c) do not hold, and let S

[26] We must also have $v \sim^* v$, and $\leq_v$ can be any relation satisfying the conditions of Definition 3.2.

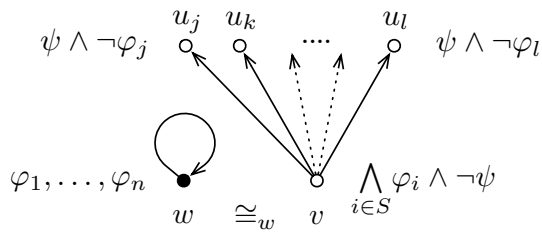


Figure 4: Construction of a countermodel for a closure principle in S-semantics.

be the largest subset of $\{1, \ldots, n\}$ such that $\psi \rightarrow \bigwedge_{i\in S}\varphi_i$ is a tautology. Then since (c) does not hold, $\bigwedge_{i\in S}\varphi_i \rightarrow \psi$ is not a tautology, so $\bigwedge_{i\in S}\varphi_i \wedge \neg\psi$ is satisfiable; and for every $j \notin S$, $\psi \wedge \neg\varphi_j$ is satisfiable by the definition of S.

We now build a countermodel $\mathcal{M}, w$ for $K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi$, as displayed in Figure 4 above. First, we make $\varphi_1, \ldots, \varphi_n$ true at w, which is possible since (b) does not hold, construct the similarity relation $\leqslant_w$ such that $w \in \mathrm{Max}_{\leqslant_w}(W)$, and construct the doxastic relation $\mathcal{B}$ such that $\mathcal{B}(w) = \{w\}$. Next, we add a world v satisfying $\bigwedge_{i\in S}\varphi_i \wedge \neg\psi$, such that $v \in \mathrm{Max}_{\leqslant_w}(W)$. Then for each $j \notin S$, we add a world u_j satisfying $\psi \wedge \neg\varphi_j$, such that $u_j \notin \mathrm{Max}_{\leqslant_w}(W)$, and we extend the relation $\mathcal{B}$ such that $\mathcal{B}(v) = \{u_j \mid j \notin S\}$. It follows that for all $j \notin S$, $B\varphi_j$ is false at v, while for all $i \in S$, φ_i is true at v by assumption. Hence for all $k \leq n$, if $B\varphi_k$ is true at v, then φ_k is true at v. Together with the fact that $B\varphi_1 \wedge \cdots \wedge B\varphi_n$ and $\varphi_1 \wedge \cdots \wedge \varphi_n$ are true at w, this implies that for all $i \leq n$, the closest$_w$ $B\varphi_i$-worlds satisfy φ_i. Hence the safety conditions are satisfied and $K\varphi_1 \wedge \cdots \wedge K\varphi_n$ is true at w. Finally, since each u_j is a ψ-world, $B\psi$ holds at v by the construction of $\mathcal{B}$. Yet ψ is false at v by construction. Hence there is a closest$_w$ $B\psi$-world that is not a ψ-world, so the safety condition is violated. We conclude that $K\psi$ is false at w in S-semantics, so $\mathcal{M}, w$ is a countermodel for $K\varphi_1 \wedge \cdots \wedge K\varphi_n \rightarrow K\psi$. □

References

M. Alspector-Kelly. Why safety doesn't save closure. *Synthese*, 2010. doi: 10.1007/s11229-010-9755-x. forthcoming.

J. van Benthem. *Modal Logic for Open Minds*. CSLI Publications, 2010.

A. Brueckner. Strategies for refuting closure for knowledge. *Analysis*, 64(4): 333–335, 2004.

B. F. Chellas. *Modal logic: an introduction*. Cambridge University Press, 1980.

S. Cohen. How to be a fallibilist. *Philosophical Perspectives*, 2:91–123, 1988.

S. Cohen. Basic knowledge and the problem of easy knowledge. *Philosophy and Phenomenological Research*, 65(2):309–329, 2002.

K. DeRose. Solving the skeptical problem. *The Philosophical Review*, 104(1): 1–52, 1995.

F. Dretske. Epistemic operators. *The Journal of Philosophy*, 67(24):1007–1023, 1970.

F. Dretske. The pragmatic dimension of knowledge. *Philosophical Studies*, 40: 363–378, 1981.

J. Hawthorne. *Knowledge and Lotteries*. Oxford University Press, 2004.

M. Heller. Relevant alternatives. *Philosophical Studies*, 55:23–40, 1989.

M. Heller. Relevant alternatives and closure. *Australasian Journal of Philosophy*, 77(2):196–208, 1999.

J. Hintikka. *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell: Cornell University Press, 1962.

J. L. Kvanvig. Closure principles. *Philosophy Compass*, 1(3):256–267, 2006.

K. Lawlor. Living without closure. *Grazer Philosophische Studien*, 69:25–49, 2005.

D. Lewis. *Counterfactuals*. Oxford: Blackwell, 1973.

D. Lewis. *On the Plurality of Worlds*. Oxford: Blackwell, 1986.

D. Lewis. Elusive knowledge. *Australasian Journal of Philosophy*, 74(4):549–567, 1996.

P. Murphy. A strategy for assessing closure. *Erkenntnis*, 65:365–383, 2006.

R. Nozick. *Philosophical Explanations*. Harvard University Press, 1981.

P. Rysiew. Motivating the relevant alternatives approach. *Canadian Journal of Philosophy*, 36(2):259–279, 2006.

E. Sosa. How to defeat opposition to Moore. *Noûs*, 33(13):141–153, 1999.

R. Stalnaker. A theory of conditionals. In N. Rescher, editor, *Studies in Logical Theory*, volume 2, pages 98–112. Oxford: Basil Blackwell, 1968.

R. Stalnaker. The problem of logical omniscience I. *Synthese*, 89:425–440, 1991.

G. Stine. Skepticism, relevant alternatives, and deductive closure. *Philosophical Studies*, 29:249–261, 1976.

J. Vogel. Subjunctivitis. *Philosophical Studies*, 134:73–88, 2007.

T. Williamson. Interview. In V. Hendricks and O. Roy, editors, *Epistemic Logic: 5 Questions*, pages 249–261. Automatic Press, 2010.

Design as Imagining Future Knowledge, a Formal Account

Lex Hendriks and Akin Kazakci

ILLC University of Amsterdam, CGS Mines ParisTech
a.hendriks@uva.nl, akin.kazakci@mines-paristech.fr

Abstract

Design, as in designing artifacts like cars or computer programs, is one of those aspects of rational agency hardly even mentioned in traditional logical theory. As an engineering discipline, design depends on a mix of factual knowledge, experimenting and imagination, but obviously also involves reasoning. Here we are interested in the type of reasoning leading to new artifacts, things we may learn to know they can exist but for the moment are mere imagination. We will present a formal framework for the dynamic interplay between knowledge and imagination inspired by C-K theory Hatchuel and Weil (2003a) and discuss the possible directions for further development of a 'logic of design'.

1 Introduction

Engineering design can be studied with the aim of discovering patterns that may optimize the process of industrial innovation or the right conditions for organizing design teams. Such research focuses for example on the use of design practices, the possibilities of automated support for designers or the role of creativity in design Geis and Birkhofer (2010). In this paper we focus on the kind of *reasoning* that seems to be typical for (conceptual) design.

Inspired by the concept knowledge (C-K) design theory we regard design

as 'making sense out of fantasy' and study the dynamic interplay of knowledge and imagination in design within a formal logical framework. Building on our previous work Kazakci (2009), Hendriks and Kazakci (2010), the current work introduces formal *Design Operators*, replacing the underspecified operators of C-K theory, and sketches how *Design Scenario's* can be built with such operators to fully explain the phenomena described in C-K theory, such as the *conceptual expansion*.

The following short story may act as an informal introduction of the interplay of knowledge and imagination and what we mean by making sense out of fantasy.

For ages mankind could only dream about what is going on in space. But those dreams made people curious and this curiosity drove them to discoveries. They did not only tell stories about gods driving divine vehicles and throwing lightning bolts, but their belief in the influence of the moon and the stars on life on earth led to predicting celestial events and the discovery of patterns in the movements of the celestial bodies.

The gradually collected knowledge about the solar system made man imagine space vehicles that could drive around, say, on Mars. Such imaginations raised new questions, e.g. about the surface and the atmosphere of Mars.

Finally, the 4th of July 1997 the Mars Rover Sojourner wheeled over the Red Planet, expanding our knowledge of Mars with new data about the rocks and the atmosphere of Mars.

The process that eventually led to the construction of the Sojourner illustrates what we mean by the dynamic interplay of knowledge and imagination in engineering design.

The plan of the paper is as follows. Section 2 gives a summary of C-K theory and previous work on its formal foundations. Section 3 introduces the basic logical framework we use. Design stages and design concepts and body of knowledge are defined and some of their properties are summarized. Section 4 introduces Design Operators and then extends these to Design Scenarios. Some concluding remarks can be found in section 5.

2 A dynamic perspective on Design Reasoning

C-K theory describes design reasoning as the dynamic interplay of knowledge and concepts. Concepts are (partial) descriptions of a new object the existence

of which cannot be decided based on the current knowledge.

Each stage in the design process is defined by its current knowledge and its current concept. The design space can be considered as the product of the knowledge space and the concept space.

- Knowledge Space
  The elements in the *Knowledge Space* are sets of knowledge, representing all the knowledge available to a designer (or to a group of designers) at a given time.

- Concept Space
  The elements of the *Concept Space* are (partial) descriptions of unknown objects that may or may not be possible to exist.

Concepts can also be taken as propositions, stating the existence of something fitting the description. Such statements will neither be true nor false at the time of their creation (e.g., *'some tires are made of dust'*).

According to C-K theory, creative design starts by adding a new and unusual property to an existing concept C to form a new concept C' (e.g. *'tires for life'*). The elaboration of a concept can then be continued either by further *expansions* (tires for life are made of silicon) or by *restrictions* (that is by adding usual properties of the initial concept, e.g. tires for life are round). Such conceptual expansions or restrictions are called *partitioning* in C-K theory.

In C-K theory new concepts are formed by combinations of concepts occurring in the propositions of existing knowledge. The designer will use his or her body of knowledge K either to further partition the concept, or to attempt a validation of a given concept. This last type of operation (K-validation) corresponds to the evaluation of the feasibility of a design description (e.g. could it exist).

Often the validation of a concept will not be readily possible. In order to validate concept C, new knowledge warranting the existence conditions of such an object should be acquired. In terms of C-K theory, knowledge should be expanded (K-expansion). Such new knowledge may bring new concepts into the game, allowing for new expansions and restrictions of the design concept C.

The central proposition of C-K theory can be expressed as "design is the interaction and dual expansion of concepts and knowledge" Hatchuel and Weil (2002; 2003a; 2009).

2.1 Previous work on formal aspects of C-K theory

While there are extensive studies on modeling and theorizing about design[1], like most engineers they often describe design as a practice or even an art.

C-K theory, with its notion of design as a creative reasoning process generating new definitions and objects, is a notable exception. But despite the mathematical references and metaphors used in the presentation of C-K theory, a formal mathematical presentation of the theory has not been provided to date and it is not clear whether such an account of the full theory would be possible.

Nevertheless, some steps in formalizing C-K theory have been taken in some recent work. Hatchuel and Weil (2003b) argues that there are significant similarities between the type of reasoning described by C-K theory and forcing, a technique used in set theory for constructing alternative set theoretic models with desired properties. It is claimed that the parallel between forcing and C-K theory is an important step for design theory in general but this issue needs more formal investigation.

In a complementary approach, Kazakci et al. (2008) shows that C-K type reasoning can be implemented with much more simple formalisms. They use propositional term logic to model the basic ideas of C-K theory. They suggest a notion of "models of K space" to emphasize that different structures (or formalism) used to model knowledge will yield different conceptive power and degrees of flexibility in reasoning.

In Kazakci (2009) a first-order logical formal account of C-K theory's core notions is presented. To emphasize the constructive aspects of a design process, intuitionistic logic is used to study the interaction and expansion of concepts and knowledge, based on the definitions of the basic notions. Building on this work, Hendriks and Kazakci (2010) complements this approach in considering the core proposition of the theory, the dual expansion of concepts and knowledge, and investigating the logical implications of such a principle.

A first attempt to describe C-K theory in a dynamic logic setting can be found in Salustri (2005) where concepts are modeled as a beliefs and design steps as a moves from a state of belief towards a state of knowledge. However, the actions used in modeling design in the action logic ALX3 (see: Huang (1994)) are more abstract than those in the Design Scenarios introduced in the current

[1] E.g. Braha and Reich (2003), Maher and Gero (1990), Maimon and Braha (1996), Marples (1960), Shai and Reich (2004a;b), Suh (1990), Takeada et al. (1990), Yoshikawa (1981), Zeng (2002)

paper and do not provide an explanation of the dynamic interplay between knowledge and imagination.

2.2 Producing concepts from knowledge

One of the intriguing ideas from C-K theory is the emerging of new *concepts* from (new) knowledge. Can we generate new concepts based on existing knowledge? This seems to provide a challenge for any approach to design theory based on logic. How do we imagine our future knowledge?

From the examples in C-K theory literature we can reconstruct a simple mechanism at work here, using the language $\mathcal{L}_K$ in which our body of knowledge K is represented. If we assume $\mathcal{L}_K$ is a first order logic language enriched with a set of constants for specific individuals and predicates (the signature of $\mathcal{L}_K$), it is only natural that extending K may also extend the signature. The new part of the body of knowledge may introduce new constants (Planck's h, the star Vega-β, President Obama) and predicates (being married to, prime, being the president of).

Concepts can now be generated from knowledge by recombination of expressions used in the body of knowledge. If *'Beatrix is the queen of The Netherlands'* then 'x is the queen of y' is a phrase in $\mathcal{L}_K$. Which allows us to form an expression like: *'Planck is the queen of Vega-β'*.

If we model phrases as formulas with one free variable x (a restriction we may lift later on) this amounts to forming conjunction of existing phrases, like in $Boat(x) \wedge Flies(x)$.

Such a new phrase (if K does not imply that $\exists x.Boat(x) \wedge Flies(x)$ and neither that $\neg\exists x.Boat(x) \wedge Flies(x)$), could be used as a concept C and starting point for design.

Combined with knowledge extension this simple mechanism will supply the design process with a wealth of new concepts without any appeal to hidden creative powers.

The mechanisms above can easily be extended further, e.g. by lifting the restriction on the type of phrases used. Like in the example *'Planck is the queen of Vega-β'*, where we could start the design turning this into the question *'Wouldn't it be nice if Planck is the queen of Vega-β?'* This could 'branch' (in a series of steps) into 'absorbing very bright light to produce both energy and comfortable back ground illumination'.

2.3 Generating knowledge from concepts

A simple mechanism of knowledge expansion occurs when the designer (or the design team) is aware of the body of knowledge K, say there knowledge is a part of K called K_0. Extending this K_0 could be done be Googling the web, searching Wikipedia, asking experts etc.

A second mechanism could involve further research in the field K_0. The experts in the field might be unable to answer the questions of the design team. Further research and experimentation could be necessary. For example we might want to use carbon-epi-hexa-fluor-plexitude for the heat shield of our Vega-β-surveyor, but it is unknown what will happen if carbon-epi-hexa-fluor-plexitude is heated above 5,000 degrees Celsius.

A third mechanism may occur when we try to combine parts of theories, say the nanotechnology with the neurobiology of human brain, say, in order to use nanotechnological devices to record firing patterns of neurons. Whether such thing is possible may be a complete new subject for scientific research.

3 A logical framework

In this section, we will describe the design process as generating design stages $\langle K;\ C\rangle$, where K is some body of knowledge and C is a concept. In C-K theory one is especially interested in design stages where C is totally new for K. Here we will allow degenerated stages that can be discarded once they are seen as inconsistent or in fact not new at all (hence C turns out to be feasible already based on K).

The design process may extend a design stage in principle in infinitely many ways. One could try to imagine all 'existing' possible 'bodies of knowledge' or all 'possible concepts' and try to describe these as (a special sort of) sets, such that the operations in the design process which transform a stage $\langle K;\ C\rangle$ into a new $\langle K';\ C'\rangle$ can be defined as a special kind of (extended) 'search operations', not unlike known search algorithms (e.g. on databases or in linear programming).

Not only Ockham's Razor makes the 'existence' of such 'Knowledge Spaces' or 'Concept Spaces' suspect, simply from a pragmatic point of view, some sort of constructive reasoning in the design process seems to be attractive. Operations in the design process defined using 'mental images' of infinite collections are certainly not constructive.

Note that according to C-K theory there is no 'algorithm' or construction

that will determine the next step in a design path. Such a path can be seen to behave 'lawless' in the intuitionistic sense Kazakci (2009).

We introduce a basic logical framework that will allow us to represent in logical terms some of the ideas in C-K theory. Without assuming too much about either the expressivity of the language or the strength of the logical rules, the reader may think of $\mathcal{L}$ as the language of predicate logic and the logic is the usual classical logic (although all our results will be valid in intuitionist logic as well: we will simply not use the axiom $A \vee A$ or the rule $A \vdash A$ of classical logic). In a constructivist type of logic, like intuitionist logic, a proof of C from K is a construction, one that under a certain conditions can be used as a recipe to construct an instance of $C(x)$ based on the constructions that exist according to K - which is a suitable characteristic for modeling the design endeavor.

3.1 Design stages and design space

We will model the body of knowledge as a finite set of formula (in the language of first order predicate logic). Such a finite theory K will always be 'partial knowledge' and hence extendable. A 'concept' or concept description, will be a formula C in the same language.

Definition 3.1. A *design stage* is a pair $s = \langle K;\ C\rangle$, where K a finite set of sentences, called the *body-of-knowledge* of s and C a sentence, called the *design concept* of s.

A design stage $\langle K;\ C\rangle$ is called *consistent* if $K \nvdash \neg C$.

A design stage $\langle K;\ C\rangle$ is called *open (closed)* if $K \nvdash C$ ($K \vdash C$).

A design stage $s = \langle K;\ C\rangle$ is called *feasible* if s is both open and consistent.

A *design step* is a pair (s_0, s_1) where s_0 and s_1 are design stages. We will often use the notation $s_0 \Rightarrow s_1$ for the design step (s_0, s_1). Usually we will also assume that $s_0 = \langle K_0;\ C_0\rangle, s_1 = \langle K_1;\ C_1\rangle$ etc.

A design step $s_0 \Rightarrow s_1$ is called *sound* ($s_0 \overset{s}{\Rightarrow} s_1$), if s_1 is consistent and for all $A \in K_0$ it is true that $K_1 \vdash A$ (i.e. $K_1 \vdash \bigwedge K_0$) and $K_1, C_1 \vdash C_0$.

Design step s_0 *implies* s_1 ($s_0 \vdash s_1$) if $\bigwedge K_0 \to C_0 \vdash \bigwedge K_1 \to C_1$.

Design step s_0 is *equivalent* ($s_0 \equiv s_1$) with s_1 if s_0 implies s_1 and s_1 implies s_0.

In the definition above there are no constraints on the type of sentence used as the design concept. Our C does not necessarily have the form $\exists x.P_0(x) \wedge \ldots \wedge P_n(x)$, where the, atomic, P_i are the desired properties for the object that we wish to come out of the design process. It is even not required that C is an existential formula. One can imagine for example that the 'thing' we try to design is a way of transforming all x with property A into some y with relationship R between the x and the y. So $C = \forall x(A(x) \rightarrow \exists y R(x, y)$ would be a conceivable design concept.

That we assume the body-of-knowledge K to be finite is not a real restriction in practice (at any moment of time each finite group of people could only be aware of a finite number of facts). We could allow for infinite K in principle, but this would slightly complicate our formulas like in the definition of $s_0 \vdash s_1$, where for an infinite K the conjunction $\bigwedge K$ formally is not defined.

The following facts follow directly from our definitions.

Observation 1. Let $s_0 = \langle K_0;\ C_0\rangle$, $s_1 = \langle K_1;\ C_1\rangle$ and $s_2 = \langle K_2;\ C_2\rangle$ be design stages.

(1) If s_0 is consistent then K_0 is consistent (i.e. $K_0 \nvdash \bot$).

(2) If $s_0 \overset{s}{\Rightarrow} s_1$ then s_0 is consistent.

(3) If $s_0 \equiv s_1$ and s_1 (or s_0) is consistent then $s_0 \overset{s}{\Rightarrow} s_1$.

(4) If $s_0 \overset{s}{\Rightarrow} s_1$ and $s_1 \overset{s}{\Rightarrow} s_2$ then $s_0 \overset{s}{\Rightarrow} s_2$.

Our definition of sound design steps is one way of formalizing the intuitive notion of one design stage 'implying' another, found in most descriptions of design (i.e. in Hatchuel and Weil (2009) and Braha and Reich (2003)). The informal notion is sometimes used in a loose sense of 'having some reason to go from s_0 to s_1'. On other occasions the 'implication chain' is a series of stages where apparently more logic is involved. Our notion of soundness is a weak form of such a logical connection, whereas the defined implication ($s_0 \vdash s_1$) is rather strong[2]

Two special cases of sound design steps may clarify the often observed difference in direction of the 'implication' between 'refining' the specifications

[2]Another notion of (strict) implication would be $s_0 \vdash\!\sim s_1$ defined as $K_0 \vdash C_0 \ \Rightarrow\ K_1 \vdash C_1$. Implication of states implies strict implication, but not the other way around.

and 'expanding' the (structural) knowledge. Note that if $K_0 = K_1$ and $s_0 \overset{s}{\Rightarrow} s_1$, then $K_0, C_1 \vdash C_0$ and hence $s_1 \vdash s_0$. On the other hand, if $C_0 = C_1$ and $s_0 \overset{s}{\Rightarrow} s_1$, $K_1 \vdash \bigwedge K_0$ and hence $s_0 \vdash s_1$.

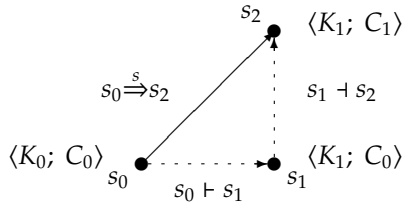


Figure 1: Splitting design steps in a K- and a C-component

As long as we confine ourselves to sound design steps that only change the design concept or only the body-of-knowledge (which theoretically we could obviously always do by splitting up steps if necessary, see figure 3.1), we could use the implication of one stage by another as a basis[3] for describing the design process (taking care each time using the right direction of the 'implication' between the states). In a more realistic model of design steps, where both parts of the state can change in one single step, the implication between states (e.g. in the way defined above) becomes awkward to deal with.

Simple examples of sound steps are[4]:

- (Adding knowledge) $\langle K;\ C\rangle \Rightarrow \langle K, A;\ C\rangle$
- (Adding properties) $\langle K;\ \exists x C(x)\rangle \Rightarrow \langle K;\ \exists x(C(x) \wedge P(x))\rangle$
- (Introducing a definition) $\langle K;\ C \wedge D\rangle \Rightarrow \langle K, P \leftrightarrow C \wedge D;\ P\rangle$

An example of $s_0 \vdash s_1$ where $s_0 \Rightarrow s_1$ is not sound would be $s_0 = \langle K;\ C\rangle$ and $s_1 = \langle K;\ C \vee D\rangle$. If $K \nvdash D \rightarrow C$ then K together with $C \vee D$ does not imply C.

As can be concluded from the examples above, our formalism does not require the design concept to be of the form $\exists x C(x)$ as we will assume in the following section on Design Scenarios.

The main theorem on sound design steps shows that a sequence of sound steps is itself again a sound step.

[3] This would result in a more restrictive relationship than soundness.

[4] Assuming the results are consistent.

Theorem 1. *Let $s_n = \langle K_n;\ C_n\rangle$ be a closed design stage reached after n sound design steps from $s_0 = \langle K_0;\ C_0\rangle$ (so $K_n \vdash C_n$) then $K_n \vdash C_0$.*

Proof. From the fact 1.3 we can conclude that $s_0 \overset{s}{\Rightarrow} s_n$. Hence, by definition, $K_n, C_n \vdash C_0$. As $K_n \vdash C_n$ by the Cut-rule it follows that $K_n \vdash C_0$. ⊣

Interestingly the notion of sound design steps makes perfectly sense to practitioners in engineering design, e.g. if safety requirements are constraining the design, but the notion seems not to have been picked up yet in design theory.

4 Design Scenarios

In C-K theory design operations are made within or between 'K-space' and 'C-space'. Translated into our formalism we get:

Create	$K \longrightarrow \langle K;\ C\rangle$
	Forming the first concept from properties of K.
Refine	$\langle K;\ C\rangle \longrightarrow \langle K;\ C'\rangle$
	Adding a property from K to C.
Enhance	$\langle K;\ C\rangle \longrightarrow \langle K';\ C\rangle$
	Using properties from C to find additional knowledge.
Expand	$\langle K;\ C\rangle \longrightarrow \langle K';\ C'\rangle$
	Combining Refine and Enhance.
Validate	$\langle K';\ C\rangle \longrightarrow \langle K;\ C\rangle$
	Adding knowledge about the existence of C to K
	This may or may not add C or $\neg C$ to K.

Publications on C-K theory treat concepts as sets of properties, as (partial) definitions of an artifact and as a proposition stating the existence of such an artifact. These statements are obviously related: if $\{\psi_1, \ldots \psi_n\}$ is a set of properties of the artifact, $\varphi(x) = \psi_1(x) \wedge \ldots \psi_n(x)$ can be regarded as a description or partial definition, whereas $\exists x \varphi(x)$ is the statement of its existence Kazakci (2009).

Here we will start with a similar formalism, introducing properties, sets of such properties and some operator $\exists$ to bridge the gap between sets of properties and the propositions C introduced before. This leads to a description of the

Design Scenarios that is closely related to the exposition of design operators e.g. in Hatchuel and Weil (2002).

Definition 4.1. Let S be a set of properties.
$\exists S$ will be the proposition claiming the existence of an object with all the properties in S.
For the design stage $\langle K;\ \exists S\rangle$ we will also use $(K;\ S)$. The other way around $\langle(K; S)\rangle$ by definition is the design stage $\langle K;\ \exists S\rangle$.

A further analysis of the design operations listed above shows that apparently also other basic extra-logical operations on (sets) of properties and propositions, like Query, Test and operators to extract properties from formulas, will play a role in describing the dynamic interplay between knowledge and imagination.

Definition 4.2. Let A be a formula and P a predicate occurring in A, than both $P(\vec{x})$ and $\neg P(\vec{x})$ are *properties* of A.
Let A be a formula and S a set, then the basic design operations are defined as:

$\neg$	For property P, $\neg P$ is property: the complement of P.
$.Prop$	$A.Prop$ is the set of properties of A.
	If S a set of formulas, $S.Prop = \bigcup\{A.Prop \mid A \in S\}$.
!	$!S$ is a subset of S.
?	If S is a set of properties, $?S$ is a sentence (obtained by a query).
test	If A is a formula $test(A)$ is a formula.

In our extended language we will allow the use of set operations, relations (e.g. $\cup, \cap, \in, \subseteq, =, \neq$) and the constant $\emptyset$ (for the empty set).

How exactly selection, e.g. of a subset of properties is made, or how the query ? or the *test* are performed we keep for now formally undefined. Several options may be investigated within (and by expanding) our formal framework. Our definition of the complement of a property does not rule out that $\neg\neg P$ is the same property as P, nor does it requires so. Usually one surely would expect $\neg\neg\neg P$ to equal $\neg P$, but for the moment we will not fix the rules of the formalism at this level of detail.

For selection one may think of a random choice. The query may be a Google-search with the English names of the properties, out of the result of which somehow some piece of knowledge is selected and translated in the language $\mathcal{L}_{K'}$, where K' may be an extension of K containing some new properties.

4.1 Scenarios

We are now ready explain what we consider the core of C-K theory, the *dual expansion of concept and knowledge*, in terms of our basic operations. Consider the following scenario:

$$\begin{array}{lll} \langle expand\rangle(K;\ S) & := \{ & P := !S; \\ & & K_P := ?P; \\ & & Q := !(K \cup K_P).Prop; \\ & & K_Q := ?Q; \\ & & R := !K_Q.Prop; \\ & & \textbf{return}\ (K \cup K_Q \cup K_P;\ S \cup R)\} \end{array}$$

This scenario can be read as: Based on some properties of S we expand our knowledge to K_P. We add this new knowledge to our existing knowledge, and choose a new set of properties Q (which may contain new properties introduced by K_P). A query fed by Q leads to (possibly new) knowledge K_Q. The set of properties for the design concept is finally extended with a subset of the properties from this K_Q whereas the body of knowledge meanwhile is extended with both K_P and K_Q.

Assuming we started with $K \nvdash \exists S$, it now might be the case that $K, K_P, K_Q \vdash \exists S$, or even $K, K_P, K_Q \vdash \exists(S \cup R)$.

Based on the definition of the $\langle expand\rangle$ operator it is not difficult to prove, within the logical framework sketched above, the following theorem.

Theorem 2. *If* $s_0 = \langle(K; S)\rangle$, $s_1 = \langle expand(K; S)\rangle$ *and* s_1 *is consistent, then* $s_0 \overset{S}{\Rightarrow} s_1$.

The scenario above kept close to the description of expansion in C-K theory, as in Hatchuel and Weil (2002). We also restricted our notion of properties to atomic formulae, but in fact the framework sketched can easily be expanded.

We also treated only *expand*, the 'show case' of C-K theory, but the other operations of C-K theory can be treated similar. For example it is reasonable to assume $!\emptyset = \emptyset$ and $?\emptyset = \top$. These assumptions would imply that one can define:

$$\langle create\rangle(K) := \langle expand\rangle(K;\ \emptyset)$$

This definition would further simplify the C-K framework of design operators (and proves *create* can be regarded as a sound design step).

Exploring the framework in yet an other direction, it seems natural to become more specific about the basic operations of the Design Scenarios and for

example tie the selection operator to the preferences of the designers (or their customers).

The real challenge however is to translate the formal Design Scenarios into a framework of dynamic logic, e.g. in the spirit of van Benthem (2011).

5 Conclusions

Our approach in this paper has been a mix of formal reasoning and informal analysis of engineering design, especially as described by C-K theory. The result is a formal framework with some 'dynamic' features for which we hope future research will provide a well defined semantics.

Our framework clarifies and explains many underspecified notions in C-K theory, including the central notion of dual expansion and introduced an interesting new notion of sound design steps. The application of a logical approach to design has been further exploited in Kazakci and Hendriks (2011) where we introduce Design Tableaux, based on Beth's semantical tableaux, to study the Design Scenarios presented in this paper from a slightly different angel.

Most important our approach allows further formal investigation of design reasoning, for example, as pointed out, applying notions and techniques from dynamic logic. This could especially be useful to develop a formal semantics for our logic based approach to conceptual design.

References

D. Braha and Y. Reich. Topological structures for modeling engineering design processes. *Research in Engineering Design*, 14:185–199, 2003.

C. Geis and H. Birkhofer. Classification and synthesis of design theories. In Marjanovic et al. (2010), pages 39–48.

A. Hatchuel and B. Weil. C-K theory: Notions and applications of a unified design theory. In *International Conference on the Sciences of Design, in honor of Herman Simon*, INSA Lyon, France, March 2002.

A. Hatchuel and B. Weil. A new approach of innovative design: an introduction to C-K design theory. In *Proceedings of the International Conference on Engineering Design (ICED03)*, 2003a.

A. Hatchuel and B. Weil. Design as forcing: Deepening the foundations of C-K theory. In J.-C. Bocquet, editor, *Proceedings of the International Conference on Engineering Design 2007*, page 14, Paris, France, August 2003b.

A. Hatchuel and B. Weil. C-K design theory: an advanced formulation. *Research in Engineering Design*, 19:181–192, 2009.

L. Hendriks and A. Kazakci. A formal account of the dual expansion of concepts and knowledge in C-K theory. In Marjanovic et al. (2010), pages 49–58.

Z. Huang. *Logics for Agents with Bounded Rationality*. PhD thesis, Institute for logic, Language and Computation, Universiteit van Amsterdam, Amsterdam, The Netherlands, December 1994. ILLC Dissertation series DS-1994-10.

A. Kazakci. A formalization of C-K design theory based on intuitionistic logic. In A. Chakrabarti, editor, *Proceedings of the International Conference on Research into Design ICORD09*, pages 499–507, Bangladore, India, January 2009.

A. Kazakci and L. Hendriks. A method for designreasoning using logic: from semantic tableaux to design tableaux. In J.-C. Bocquet, editor, *Proceedings of the International Conference on Engineering Design (ICED11), to appear*, Stanford, California, August 2011.

A. Kazakci, A. Hatchuel, and B. Weil. A model of CK design theory based on based on a term logic: a formal background for a class of design assistants. In D. Marjanovic, M. Storga, N. Pavkovic, and N. Bojcetic, editors, *Proceedings of the 10th International Design Conference DESIGN 2008*, pages 43–52, Dubrovnik, Croatia, May 2008.

M. Maher and J. Gero. Theoretical requirements for creative design by analogy. In P. A. Fitzhorn, editor, *Proceedings of the first International Workshop on Formal Methods in Engineering Design, Manufacturing, and Assembly*, pages 19–27, Colorado Springs, Colorado, January 1990.

O. Maimon and D. Braha. A mathematical theory of design. *International Journal of General Systems*, 27:275–318, 1996.

D. Marjanovic, M. Storga, N. Pavkovic, and N. Bojcetic, editors. *Proceedings of the 11th International Design Conference DESIGN 2010*, Dubrovnik, Croatia, May 2010.

D. Marples. The decisions of engineering design. *Journal of the Institute of Engineering Designers*, pages 181–192, December 1960.

P. A. Salustri. Representing CK theory with an action logic. In A. Samuel and W. Lewis, editors, *Proceedings of the 15th International Conference on Engineering Design 2005*, Melbourn, Australia, August 2005.

O. Shai and Y. Reich. Infused design: I Theory. *Research in Engineering Design*, 15:93–107, 2004a.

O. Shai and Y. Reich. Infused design: II Practice. *Research in Engineering Design*, 15:108–121, 2004b.

N. P. Suh. *The principles of design*. Oxford University Press, New York, 1990.

H. Takeada, P. Veerkamp, T. Tomiyama, and H. Yoshikawa. Modeling design processes. *AI Magazine*, 11:37–48, 1990.

J. van Benthem. *Logical Dynamics of Information and Interaction*. Cambridge University Press, Cambridge, UK, 2011.

H. Yoshikawa. General design theory and a CAD system. In T. Sata and E. Waterman, editors, *Man-Machine Communication in CAD/CAM*, pages 35–38. North-Holland, 1981.

Y. Zeng. Axiomatic theory of design modeling. *Journal of Integrated Design and Process Science*, 6:1–28, August 2002. ISSN 1092-0617.

Binary Aggregation with Integrity Constraints

Umberto Grandi and Ulle Endriss

Institute for Logic, Language and Computation, University of Amsterdam
u.grandi@uva.nl, ulle.endriss@uva.nl

Abstract

We consider problems where several individuals each need to make a yes/no choice regarding a number of issues and these choices then need to be aggregated into a collective choice. We describe rationality assumptions in terms of a propositional formula and we explore the question of whether or not a given aggregation procedure will *lift* the rationality assumptions from the individual to the collective level. For various fragments of propositional logic, we provide an axiomatic characterisation of the class of aggregation procedures that will lift all rationality assumptions expressible in this fragment. We also show how several classical frameworks of Social Choice Theory, particularly preference and judgment aggregation, can be viewed as binary aggregation problems by designing suitable integrity constraints.

1 Introduction

In recent years, the AI community has dedicated more and more attention to the study of methods coming from Social Choice Theory (SCT). The reasons for this focus are clear: SCT provides tools for the analysis of collective choices of groups of agents, and as such is of immediate relevance to the study of multiagent systems. At the same time, studies in AI have led to a new and broadened perspective on classical results in SCT, e.g., via the use of knowledge

representation languages for modelling preferences in social choice problems or via the complexity-theoretic analysis of the implementation of social choice rules. Particularly close to the interests of AI is the problem of social choice in combinatorial domains (Chevaleyre et al. 2008), where the space of choices the individuals have to make has a combinatorial structure.

Many of the questions studied in SCT arise from the observation of paradoxes, such as the Condorcet Paradox in preference aggregation (Gaertner 2006) or the Doctrinal Paradox in judgment aggregation (List and Puppe 2009). One of the scopes of this paper is to show how these can all be viewed as instances of a general definition of *paradox*, and to do so we translate classical frameworks for SCT into a canonical (and more easily implementable) one. This framework is *binary aggregation with integrity constraints*, introduced and studied in our previous works (Grandi and Endriss 2010; 2011), building on research initiated by Wilson (1975) and more recently developed by Dokow and Holzman (2010).

Dokow and Holzman (2010) characterise domains of aggregation over which every independent and unanimous procedure is dictatorial. This is a good example for the use of the axiomatic method in economic theory: the aim is to identify *the* appropriate set of axioms (e.g., to model real-word economies, specific moral ideals, etc.) and then to prove a characterisation (or impossibility) result for those axioms. AI suggests an alternative approach: with every new application the principles underlying a system may change; so we may be more interested in devising languages for expressing a range of different axioms rather than identifying the "right" set of axioms; and we may be more interested in developing methods that will help us to understand the dynamics of a range of different social choice scenarios rather than in technical results for a specific such scenario.

For this purpose we separate two parameters in the framework of binary aggregation. On the one hand, we introduce a propositional language to define the domain of aggregation by expressing a rationality assumption common to all individuals. On the other, we state a list of axioms to classify aggregation procedures over these domains. We call an aggregation procedure *collectively rational* with respect to a language if whenever all individuals submit ballots satisfying a formula in the language, so does the outcome of aggregation. We characterise, for several simple fragments of the language of propositional logic, the associated class of collectively rational procedures as the set of procedures satisfying a certain set of axioms.

We then turn to the study of classical frameworks of SCT as instances of binary aggregation with integrity constraints. We show how characterisation results proved in binary aggregation can be used to derive a new impossibility

theorem in preference aggregation, a variant of Arrow's Theorem, by identifying a clash between the syntactic shape of the integrity constraints defining the framework of preference aggregation and a number of axiomatic postulates. In a similar fashion, we are able to translate problems in judgment aggregation into binary aggregation problems with a specific integrity constraint, and we identify a syntactic analogue of classical agenda properties guaranteeing consistent aggregation.

The paper is organised as follows: Section 2 presents the framework of binary aggregation with integrity constraints. In Section 3 we prove several characterisation results relating axiomatic requirements and collective rationality. Section 4 and Section 5, respectively, deal with the translation of preference and judgment aggregation to binary aggregation, and Section 6 concludes.

2 Binary Aggregation with Integrity Constraints

In this section we introduce the framework of binary aggregation with integrity constraints, based on work by Wilson (1975) and Dokow and Holzman (2010). We introduce two crucial definitions: a new definition of the notion of *paradox* and the definition of *collective rationality*. We conclude by stating classical axioms for aggregation procedures adapted to the framework of binary aggregation.

2.1 Terminology and Notation

Let $\mathcal{I} = \{1, \dots, m\}$ be a finite set of *issues*, and let $\mathcal{D} = D_1 \times \cdots \times D_m$ be a boolean combinatorial *domain*, i.e., $|D_i| = 2$ for all $i \in \mathcal{I}$ (we assume $D_i = \{0, 1\}$). Let $PS = \{p_1, \dots, p_m\}$ be a set of propositional symbols, one for each issue, and let $\mathcal{L}_{PS}$ be the corresponding propositional language. For any $\varphi \in \mathcal{L}_{PS}$, let $\mathrm{Mod}(\varphi)$ be the set of *models* that satisfy φ. For example, $\mathrm{Mod}(p_1 \wedge \neg p_2) = \{(1, 0, 0), (1, 0, 1)\}$ if $PS = \{p_1, p_2, p_3\}$. We call *integrity constraint* any formula $\mathrm{IC} \in \mathcal{L}_{PS}$. Any such formula defines a *domain of aggregation* $X := \mathrm{Mod}(\mathrm{IC})$.

Integrity constraints can be used to define what tuples in $\mathcal{D}$ we consider *rational* choices. For example, as we shall see in Section 4, $\mathcal{D}$ might be used to encode a binary relation representing preferences, in which case we may want to declare only those elements of $\mathcal{D}$ rational that correspond to relations that are transitive. We shall therefore use the terms "integrity constraints" and "rationality assumptions" interchangeably.

Let $\mathcal{N} = \{1, \ldots, n\}$ be a finite set of *individuals*. A *ballot* B is an element of $\mathcal{D}$ (i.e., an assignment to the variables $p_1, \ldots, p_m$); and a *rational ballot* B is an element of $\mathcal{D}$ that satisfies the integrity constraint, i.e., an element of Mod(IC). A *profile* $\mathbf{B}$ is a vector of (rational) ballots, one for each individual in $\mathcal{N}$. We write b_j for the jth element of a ballot B, and $b_{i,j}$ for the jth element of ballot B_i within a profile $\mathbf{B} = (B_1, \ldots, B_n)$. An *aggregation procedure* is a function $F : \mathcal{D}^{\mathcal{N}} \to \mathcal{D}$, mapping each profile to an element of $\mathcal{D}$. $F(\mathbf{B})_j$ denotes the result of the aggregation on issue j.

2.2 Paradoxes and Collective Rationality

Consider the following example: Let IC $= \neg(p_1 \wedge p_2 \wedge p_3)$ and suppose there are three individuals, choosing $(1, 1, 0)$, $(1, 0, 1)$ and $(0, 1, 1)$, respectively, i.e., their choices are rational (they all satisfy IC). If we use issue-wise majority (accepting p_i if a majority of individuals do) to aggregate their choices, however, we obtain $(1, 1, 1)$, which fails to be rational. This kind of observation is often referred to as a paradox.

We now give a general definition of paradoxical behaviour of an aggregation procedure in terms of the violation of certain rationality assumptions:

Definition 2.1. A paradox is a triple $(F, \mathbf{B}, \text{IC})$, where $F : \mathcal{D}^{\mathcal{N}} \to \mathcal{D}$ is an aggregation procedure, $\mathbf{B}$ is a profile in $\mathcal{D}^{\mathcal{N}}$, IC $\in \mathcal{L}_{PS}$, and $B_i \models \text{IC}$ for all $i \in \mathcal{N}$ but $F(\mathbf{B}) \not\models \text{IC}$.

As we shall see in the following sections, various classical paradoxes in SCT are instances of this definition. A closely related notion is that of collective rationality:

Definition 2.2. Given an integrity constraint IC $\in \mathcal{L}_{PS}$, an aggregation procedure $F : \mathcal{D}^{\mathcal{N}} \to \mathcal{D}$ is called collectively rational (CR) for IC, if for all rational profiles $\mathbf{B} \in \text{Mod(IC)}^{\mathcal{N}}$ we have that $F(\mathbf{B}) \in \text{Mod(IC)}$.

Thus, F is CR if it can *lift* the rationality assumptions given by IC from the individual to the collective level. An aggregation procedure that is CR with respect to IC cannot generate a paradox with IC as integrity constraint.

2.3 Axiomatic Method

In social choice theory, aggregation procedures are studied using the axiomatic method. Axioms are used to express desirable properties of a procedure. In

this section, we adapt the most important axioms familiar from standard social choice theory, and more specifically from judgment aggregation (List and Puppe 2009) and binary aggregation theory (Dokow and Holzman 2010), to our setting. We start with four common axioms:

Unanimity (U): For any profile $\mathbf{B} \in X^N$ and any $x \in \{0,1\}$, if $b_{i,j} = x$ for all $i \in N$, then $F(\mathbf{B})_j = x$.

Anonymity (A): For any profile $\mathbf{B} \in X^{\mathcal{N}}$ and any permutation $\sigma : \mathcal{N} \to \mathcal{N}$, we have that $F(B_1, \ldots, B_n) = F(B_{\sigma(1)}, \ldots, B_{\sigma(n)})$.

Issue-Neutrality ($\mathrm{N}^{\mathcal{I}}$): For any two issues $j, j' \in \mathcal{I}$ and any profile $\mathbf{B} \in X^{\mathcal{N}}$, if for all $i \in \mathcal{N}$ we have that $b_{i,j} = b_{i,j'}$, then $F(\mathbf{B})_j = F(\mathbf{B})_{j'}$.

Independence (I): For any issue $j \in \mathcal{I}$ and profiles $\mathbf{B}, \mathbf{B}' \in X^{\mathcal{N}}$, if $b_{i,j} = b'_{i,j}$ for all $i \in \mathcal{N}$, then $F(\mathbf{B})_j = F(\mathbf{B}')_j$.

Unanimity postulates that, if all individuals agree on issue j, then the aggregation procedure should implement that choice for j. Anonymity requires the procedure to be symmetric with respect to individuals. Issue-neutrality (a variant of the standard axiom of neutrality introduced in judgment aggregation) asks that the procedure be symmetric with respect to issues. Finally, independence requires the outcome of aggregation on a certain issue j to depend only on the individual choices regarding that issue. Combining independence with issue-neutrality, we get the axiom of systematicity (S) = (I) + ($\mathrm{N}^{\mathcal{I}}$).

It is important to remark that all axioms are domain-dependent. For instance, many aggregation procedures, such as the majority rule, are independent over the full combinatorial domain $\mathcal{D}$, while others, such as the one presented in the next example, are not. With two issues, let IC = $(p_2 \to p_1)$ and let F be equal to the majority rule on the first issue, and accept the second issue only if the first one was accepted and the second one has the support of a majority of the individuals. This procedure is not independent on the full domain, but it is easy to see that it satisfies independence when restricted to $X^N = \mathrm{Mod}(\mathrm{IC})^N$.

As a generalisation of the axiom of neutrality introduced by May (1952), we introduce the following:

Domain-Neutrality ($\mathrm{N}^{\mathcal{D}}$): For any two issues $j, j' \in \mathcal{I}$ and any profile $\mathbf{B} \in X^{\mathcal{N}}$, if $b_{i,j} = 1 - b_{i,j'}$ for all $i \in \mathcal{N}$, then $F(\mathbf{B})_j = 1 - F(\mathbf{B})_{j'}$.

The two notions of neutrality are uncorrelated but dual: issue-neutrality requires the outcome on two issues to be the same if all individuals agree on these issues; domain-neutrality requires it to be reversed if all the individuals make opposed choices on the two issues.

The following axiom of monotonicity is often called *positive responsiveness*, and is formulated as an (inter-profile) axiom for independent aggregation procedures:[1]

I-Monotonicity (M): For any issue $j \in \mathcal{I}$ and profiles $\mathbf{B}=(B_1..B_i..B_n)$ and $\mathbf{B'}=(B_1..B'_i..B_n)$ in $X^{\mathcal{N}}$, if $b_{i,j}=0$ and $b'_{i,j}=1$, then $F(\mathbf{B})_j = 1$ entails $F(\mathbf{B'})_j = 1$.

Every set of axioms identifies a class of aggregation procedures that satisfy these properties. A characterisation in mathematical terms can be obtained for some classes. One example is the class of *quota rules* $\mathcal{QR}$ introduced by Dietrich and List (2007): an aggregation procedure F for n individuals is a quota rule if for every issue j there exists a quota $0 \leqslant q_j \leqslant n+1$ such that, if we denote by $N_j^{\mathbf{B}} = |\{i \mid b_{i,j}=1\}|$, then $F(\mathbf{B})_j=1$ if and only if $N_j^{\mathbf{B}} \geqslant q_j$. The following representation result holds:

Proposition 1 (Dietrich and List, 2007)**.** *An aggregation procedure F satisfies* A, I, *and* M^{I} *if and only if it is a quota rule.*

A quota rule is called *uniform* if the quota is the same for all issues. By adding the axiom of issue-neutrality to Proposition 1 we get an axiomatisation of this class. The uniform quota rule with $q_j = \lceil \frac{n}{2} \rceil$ for all issues j is the *majority rule*. If n is odd, then the majority rule satisfies all of the axioms listed above—but, as we have seen, it is not CR even for simple integrity constraints such as $\neg(p_1 \wedge p_2 \wedge p_3)$. It is interesting to link these results with May's Theorem (1952) on the axiomatic characterisation of the majority rule in voting. We can obtain a more general version of his result (which deals with the case of a single issue) by adding the axiom of domain-neutrality: this forces the quota to treat $N_j^{\mathbf{B}}$ and $n - N_j^{\mathbf{B}}$ symmetrically, and thus the only possibility is to fix the quota as the majority of the individuals.

[1] A variant of this axiom for issue-neutral aggregators has been defined in previous work (Endriss et al. 2010).

3 Lifting Individual Rationality

We now want to establish connections between aggregation procedures characterised in terms of axioms and aggregation procedures characterised in terms of languages for integrity constraints for which they are collectively rational. To this end, we first define the class of procedures that can lift the integrity constraints belonging to a given language $\mathcal{L}$ (recall Definition 2.2).

Definition 3.1. For any language $\mathcal{L} \subseteq \mathcal{L}_{PS}$, define the class $\mathcal{CR}[\mathcal{L}]$ of aggregation procedures that lift $\mathcal{L}$:

$$\mathcal{CR}[\mathcal{L}] \quad = \quad \{F : \mathcal{D}^N \to \mathcal{D} \mid F \text{ is CR for all IC} \in \mathcal{L}\}$$

Next, we establish some basic properties of $\mathcal{CR}[\mathcal{L}]$. In our framework, we have made the assumption of IC being a single formula (rather than a set of formulas); we now provide a formal underpinning for this choice. For any $\mathcal{L} \subseteq \mathcal{L}_{PS}$, let $\mathcal{L}^\wedge$ be the language of conjunctions of formulas in $\mathcal{L}$.

Lemma 1. $\mathcal{CR}[\mathcal{L}^\wedge] = \mathcal{CR}[\mathcal{L}]$ *for all* $\mathcal{L} \subseteq \mathcal{L}_{PS}$.

Proof. $\mathcal{CR}[\mathcal{L}^\wedge]$ is clearly included in $\mathcal{CR}[\mathcal{L}]$, since $\mathcal{L} \subseteq \mathcal{L}^\wedge$. It remains to be shown that, if an aggregation procedure F lifts every constraint in $\mathcal{L}$, then it lifts any conjunction of formulas in $\mathcal{L}$. Let $\bigwedge_k \mathrm{IC}_k$ be a conjunction of formulas in $\mathcal{L}$, and let $\mathbf{B} \in \mathrm{Mod}(\bigwedge_k \mathrm{IC}_k)^N$ be a profile satisfying this integrity constraint. Since $\mathrm{Mod}(\bigwedge_k \mathrm{IC}_k) = \bigcap_k \mathrm{Mod}(\mathrm{IC}_k)$, we have that $\mathbf{B} \in \mathrm{Mod}(\mathrm{IC}_k)$ for every k. Thus, if $F \in \mathcal{CR}[\mathcal{L}]$, then $F(\mathbf{B}) \in \mathrm{Mod}(\mathrm{IC}_k)$ for every k. Therefore, F will also be in $\mathrm{Mod}(\bigwedge_k \mathrm{IC}_k)$, and this concludes the proof. □

In particular, we have that $\mathcal{CR}[\mathit{cubes}] = \mathcal{CR}[\mathit{literals}]$ and $\mathcal{CR}[\mathit{clauses}] = \mathcal{CR}[\mathcal{L}_{PS}]$. The latter holds, because for every propositional formula there is an equivalent formula in conjunctive normal form (CNF).

The following lemma is an immediate consequence of our definitions:

Lemma 2. $\mathcal{CR}[\mathcal{L}_1 \cup \mathcal{L}_2] = \mathcal{CR}[\mathcal{L}_1] \cap \mathcal{CR}[\mathcal{L}_2]$ *for all* $\mathcal{L}_1, \mathcal{L}_2 \subseteq \mathcal{L}_{PS}$.

Next we introduce notation for defining classes of aggregation procedures in terms of axioms. As mentioned earlier, a particular axiom may be satisfied on a subdomain of interest, but not on the full domain. Here, our subdomains of interests are subdomains that correspond to rational ballots for a given constraint. We therefore need to be able to speak about the procedures that satisfy an axiom on the subdomain $\mathrm{Mod}(\mathrm{IC})^N$ induced by a given integrity constraint IC.

Let $F_{\restriction \mathrm{Mod(IC)}^N}$ denote the restriction of the aggregation procedure F to the subdomain $\mathrm{Mod(IC)}^N$.

Definition 3.2. An aggregation procedure F satisfies a set of axioms AX wrt. a language $\mathcal{L} \subseteq \mathcal{L}_{PS}$, if for all constraints $\mathrm{IC} \in \mathcal{L}$ the restriction $F_{\restriction \mathrm{Mod(IC)}^N}$ satisfies the axioms in AX. This defines the following class:

$$\mathcal{F}_{\mathcal{L}}[\mathrm{AX}] \;=\; \{F : \mathcal{D}^N \to \mathcal{D} \mid F_{\restriction \mathrm{Mod(IC)}^N} \text{ sat. AX for all } \mathrm{IC} \in \mathcal{L}\}$$

We write $\mathcal{F}[\mathrm{AX}]$ as a shorthand for $\mathcal{F}_{\{\top\}}[\mathrm{AX}]$, the class of procedures that satisfy AX over the *full* domain $\mathcal{D}$. It is easy to see that the following lemma holds:

Lemma 3. $\mathcal{F}[\mathrm{AX}] \subseteq \mathcal{F}_{\mathcal{L}}[\mathrm{AX}]$ *for all* $\mathcal{L} \subseteq \mathcal{L}_{PS}$.

We shall now seek to obtain results that link the two kinds of classes defined, i.e., results of the form

$$\mathcal{CR}[\mathcal{L}] \quad = \quad \mathcal{F}_{\mathcal{L}}[\mathrm{AX}],$$

for certain languages $\mathcal{L}$ and certain axioms AX.

3.1 Characterisation Results

Our first characterisation result shows that the aggregation procedures that can lift all rationality constraints expressible in terms of a conjunction of literals (a cube) is precisely the class of unanimous procedures:

Proposition 2. $\mathcal{CR}[\mathit{cubes}] = \mathcal{F}_{\mathit{cubes}}[\mathrm{U}]$.

Proof. One direction is easy: If X is a domain defined by a cube, then every individual must agree on every literal in the conjunction, and, by unanimity, so will the collective. For the other direction, suppose that $F \in \mathcal{CR}[\mathit{cubes}]$. Fix $j \in \mathcal{I}$. Pick a profile $\mathbf{B} \in \mathcal{D}^n$ such that $b_{i,j} = 1$ (or 0) for all $i \in N$. That is, $\mathbf{B} \in \mathrm{Mod}(p_j)^N$ (or $\neg p_j$, respectively). Since F is collectively rational on every domain defined by a cube (and this includes literals), it must be the case that $F(\mathbf{B})_j = 1$ (or 0, respectively), proving the unanimity of the aggregator. □

Observe that, as $\mathcal{F}_{\mathit{cubes}}[\mathrm{U}] = \mathcal{F}[\mathrm{U}]$, the explicit mentioning of cubes on the righthand side of Proposition 2 is not needed; we chose this form of presentation for uniformity with later results on other axioms.[2] By Lemma 1, we also get $\mathcal{CR}[\mathit{literals}] = \mathcal{F}_{\mathit{literals}}[\mathrm{U}]$ (it is easy to see that $\mathcal{F}_{\mathit{literals}}[\mathrm{U}] = \mathcal{F}_{\mathit{cubes}}[\mathrm{U}]$).

[2] The same remark applies to Propositions 3 and 4 below.

Let $\mathcal{L}_{\leftrightarrow}$ be the language of bi-implications of positive literals: $\mathcal{L}_{\leftrightarrow} = \{p_j \leftrightarrow p_k \mid p_j, p_k \in PS\}$. This language allows us to characterise issue-neutral aggregators:

Proposition 3. $\mathcal{CR}[\mathcal{L}_{\leftrightarrow}] = \mathcal{F}_{\mathcal{L}_{\leftrightarrow}}[\mathrm{N}^I]$.

Proof. To prove the first inclusion ($\supseteq$), pick a positive bi-implication $p_j \leftrightarrow p_k$: issues j and k share the same pattern of acceptances/rejections and since the procedure is neutral over issues, we get $F(\mathbf{B})_j = F(\mathbf{B})_k$. The constraint is therefore lifted. For the other direction ($\subseteq$), suppose that a profile $\mathbf{B}$ is such that $b_{i,j} = \mathbf{B}_{i,k}$ for every $i \in N$. Then $\mathbf{B} \in \mathrm{Mod}(p_j \leftrightarrow p_k)^N$, and if F is in $\mathcal{CR}[\mathcal{L}_{\leftrightarrow}]$, then $F(\mathbf{B})_j$ must be equal to $F(\mathbf{B})_k$. Since this holds for every such $\mathbf{B}$, this proves that F is neutral over issues. □

Let $\mathcal{L}_{\nleftrightarrow}$ be the language of bi-implications of one negative and one positive literal: $\mathcal{L}_{\nleftrightarrow} = \{p_j \leftrightarrow \neg p_k \mid p_j, p_k \in PS\}$. That is, $\mathcal{L}_{\nleftrightarrow}$ is the language of XOR-formulas over pairs of positive literals. With a proof analogous to the one above we can characterise domain-neutrality:

Proposition 4. $\mathcal{CR}[\mathcal{L}_{\nleftrightarrow}] = \mathcal{F}_{\mathcal{L}_{\nleftrightarrow}}[\mathrm{N}^{\mathcal{D}}]$.

Let $\mathcal{F} = \{F : \mathcal{D}^N \to \mathcal{D}\}$ be the class of *all* aggregation procedures (for fixed $\mathcal{D}$ and N). The next result is an immediate consequence of our definitions:

Proposition 5. $\mathcal{CR}[\{\bot\}] = \mathcal{CR}[\{\top\}] = \mathcal{F}$.

Hence, by Lemma 2, $\mathcal{CR}[\mathcal{L} \cup \{\bot\}] = \mathcal{CR}[\mathcal{L}]$, which shows that unsatisfiable formulas can be omitted from languages for integrity constraints.

We now move on to characterising more specific classes of procedures. A *dictatorship* is an aggregation procedure that copies in every profile the ballot of a certain fixed individual, the dictator. The class $\mathcal{F}_{\mathcal{L}}[\mathrm{DIC}]$ is composed by all functions that are dictatorships when restricted to $\mathrm{Mod}(\mathrm{IC})^N$ for all $\mathrm{IC} \in \mathcal{L}$. Now, let us call a language $\mathcal{L} \subseteq \mathcal{L}_{PS}$ *trivial*, if it is composed only of formulas having a single model each. Clearly:

Proposition 6. *If $\mathcal{L}$ is trivial, then* $\mathcal{CR}[\mathcal{L}] = \mathcal{F}_{\mathcal{L}}[\mathrm{DIC}]$.

We propose the following definition of a class of aggregators that generalises the notion of dictatorship:

Definition 3.3. An aggregation procedure $F : \mathcal{D}^N \to \mathcal{D}$ is a generalised dictatorship, if there exists a map $g : \mathcal{D}^N \to N$ such that $F(\mathbf{B}) = \mathbf{B}_{g(\mathbf{B})}$ for every $\mathbf{B} \in \mathcal{D}^N$.

That is, a generalised dictatorship copies the ballot of a (possibly different) individual in every profile. Call this class $\mathcal{F}$[GDIC]. This class fully characterises the aggregators that can lift *any* integrity constraint:

Proposition 7. $\mathcal{CR}[\mathcal{L}_{PS}] = \mathcal{F}[\text{GDIC}]$.

Proof. Clearly, every generalised dictatorship lifts any arbitrary integrity constraint IC $\in \mathcal{L}_{PS}$. To prove the other direction, suppose that $F \notin \mathcal{F}[\text{GDIC}]$. Then there exists a profile $\mathbf{B} \in \mathcal{D}^N$ such that $F(\mathbf{B}) \neq \mathbf{B}_i$ for all $i \in N$. This means that for every i there exists an issue j_i such that $F(\mathbf{B})_{j_i} \neq \mathbf{B}_{i,j_i}$. Define now ℓ_{j_i} to be equal to p_{j_i} if $\mathbf{B}_{i,j_i} = 1$, to $\neg p_{j_i}$ otherwise. Call IC the following formula: $\bigvee_i \ell_{j_i}$. Clearly, $\mathbf{B}_i \models \text{IC}$ for every $i \in N$, i.e., $\mathbf{B}$ is a rational profile for the integrity constraint IC. Since $F(\mathbf{B}) \not\models \text{IC}$ by construction, F is not in $\mathcal{CR}[\{\text{IC}\}]$ and therefore also not in $\mathcal{CR}[\mathcal{L}_{PS}]$. □

All of the characterisation results presented thus far characterise a class of procedures determined by a *single* axiom (or apply to a very specific class of procedures) and by a *uniform* description of the language. So we might ask to what extent such results can be combined to allow us to make predictions regarding the collective rationality of procedures satisfying several such axioms, or in the case where the integrity constraints can be chosen from a more complex language. To illustrate the application of our results to such cases, suppose $\mathcal{CR}[\mathcal{L}_1] = \mathcal{F}_{\mathcal{L}_1}[\text{AX}_1]$ and $\mathcal{CR}[\mathcal{L}_2] = \mathcal{F}_{\mathcal{L}_2}[\text{AX}_2]$. Then Lemma 2 and the fact that $\mathcal{F}_{\mathcal{L}_1 \cup \mathcal{L}_2}[\text{AX}_1, \text{AX}_2] \subseteq \mathcal{F}_{\mathcal{L}_1}[\text{AX}_2] \cap \mathcal{F}_{\mathcal{L}_2}[\text{AX}_2]$ entail $\mathcal{F}_{\mathcal{L}_1 \cup \mathcal{L}_2}[\text{AX}_1, \text{AX}_2] \subseteq \mathcal{CR}[\mathcal{L}_1 \cup \mathcal{L}_2]$. (But note that the other inclusion is not always true.) Now, if we start from the language $\mathcal{L}_1 \cup \mathcal{L}_2$ or any of its sublanguages, then this shows that picking procedures from $\mathcal{F}_{\mathcal{L}_1 \cup \mathcal{L}_2}[\text{AX}_1, \text{AX}_2]$ is a sufficient condition for collective rationality. If, instead, we start from the axioms in AX_1 and AX_2, then we can infer that the procedures we obtain will lift any language $\mathcal{L} \subseteq \mathcal{L}_1 \cup \mathcal{L}_2$, since $\mathcal{F}[\text{AX}_1, \text{AX}_2] \subseteq \mathcal{F}_{\mathcal{L}_1 \cup \mathcal{L}_2}[\text{AX}_1, \text{AX}_2] \subseteq \mathcal{CR}[\mathcal{L}_1 \cup \mathcal{L}_2] \subseteq \mathcal{CR}[\mathcal{L}]$. (The first inclusion follows from Lemma 3.)

3.2 Negative Results

For two important classes of aggregators, it is not possible to obtain a characterisation result:

Proposition 8. *There is no language $\mathcal{L} \subseteq \mathcal{L}_{PS}$ such that $\mathcal{CR}[\mathcal{L}] = \mathcal{F}_{\mathcal{L}}$[I].*

Proof. We prove this proposition by constructing, for any choice of a language $\mathcal{L}$, an independent function that is not collectively rational for a certain IC $\in \mathcal{L}$.

Fix a language $\mathcal{L}$. W.l.o.g., this language will contain a falsifiable formula φ (otherwise $\mathcal{CR}[\mathcal{L}] = \mathcal{F}$ by Proposition 5 and we are done, as $\mathcal{F} \neq \mathcal{F}_{\mathcal{L}}$[I]). Choose a ballot/model $B^\star \in \mathcal{D}$ such that $B^\star \nvDash \varphi$. Then the constant function $F \equiv B^\star$ is an independent function (on the full domain) that is not collectively rational. □

Proposition 9. *There is no language $\mathcal{L} \subseteq \mathcal{L}_{PS}$ such that $\mathcal{CR}[\mathcal{L}] = \mathcal{F}_{\mathcal{L}}$[A].*

Proof. Employing a different technique than in the previous proof, we show that for every language $\mathcal{L}$ there exists a procedure that is collectively rational but not anonymous. First, in case $\mathcal{L}$ is trivial, by Proposition 6, $\mathcal{CR}[\mathcal{L}] = \mathcal{F}_{\mathcal{L}}$[DIC], which is strictly included in the class of all anonymous functions. Second, if $\mathcal{L}$ is not trivial, then a dictatorship is always collectively rational (cf. Proposition 7), and it is not anonymous since due to nontriviality there is an $\mathrm{IC} \in \mathcal{L}$ that allows for at least two different rational ballots. □

These results are coherent with the intuition that any assumption of collective rationality of an aggregator can only condition the outcome in view of a single profile at a time, without being able to express inter-profile requirements such as anonymity and independence. Similar remarks apply to the axiom of monotonicity (note that M^{I} is meaningful only in connection with I).

3.3 Quota Rules and Languages of Clauses

In view of the negative results proved above, we now focus on procedures satisfying anonymity, independence and monotonicity, and analyse the ability of procedures to lift rationality assumptions *within* that class. In previous work (Grandi and Endriss 2010) we proved several preliminary results for collective rationality of quota rules with respect to language of clauses. Here instead we present a recent result (Grandi and Endriss 2011) that gives precise bounds on quotas for languages of positive clauses.

Recall from Proposition 1 that the independent, anonymous and monotone procedures are exactly the quota rules, i.e., procedures that assign a quota q_j to every issue j such that $F(\mathbf{B})_j = 1 \Leftrightarrow |\{i \mid \mathbf{B}_{i,j} = 1\}| \geqslant q_j$. That is, in our notation, $\mathcal{QR} = \mathcal{F}[\mathrm{A,I,M}^{\mathrm{I}}]$. By Proposition 7 and Lemma 1, we know that $\mathcal{CR}$[*clauses*] is the collection of generalised dictatorships. Therefore, to obtain results for more attractive classes of procedures, we restrict attention to clauses of limited length. For $k \geqslant 1$, let *k-clauses* be the set of clauses of length $\leqslant k$, *k-pclauses* the set of positive k-clauses, i.e., disjunctions where all literals are positive.

Proposition 10. *A quota rule is CR for a k-pclause* IC *if and only if* $\sum_j q_j < n + k$, *with* j *ranging over all issues that occur in* IC *and* n *being the number of individuals, or* $q_j = 0$ *for at least one issue* j *that occurs in* IC.

Proof. Suppose IC $= p_1 \vee \cdots \vee p_k$ and call $i_1, \ldots, i_k$ the corresponding issues. Given that IC is a positive clause, the only way to generate a paradox is by rejecting all issues $i_1, \ldots, i_k$. Suppose that we can create a paradoxical profile **B**. Suppose moreover that all quotas are > 0 (for otherwise one issue is always accepted and the IC trivially lifted). Every individual ballot B_i must accept at least one issue to satisfy the integrity constraint; therefore the profile **B** contains at least n acceptances. Since $F(\mathbf{B})_j = 0$ for all $j = 1, \ldots k$, we have that the number of individuals accepting an issue j is strictly lower than q_j. As previously remarked, there are at least n acceptances on the profile **B**; hence, this is possible if and only if $n \leqslant \sum_j (q_j - 1)$. Therefore, we can construct a paradox with this IC if and only if $n + k \leqslant \sum_j q_j$, and by taking the contrapositive we obtain the statement of Proposition 10. □

4 Preference Aggregation

In this section we give a translation of the framework of preference aggregation for linear orders into binary aggregation for a particular language of integrity constraints.

The framework of *preference aggregation* (see e.g. Gaertner 2006) considers a finite set of individuals $\mathcal{N}$ expressing preferences over a finite set of alternatives $\mathcal{X}$. A preference relation is represented by a binary relation P over $\mathcal{X}$. Here, we shall assume that P is a linear order, i.e., an antisymmetric, transitive and complete binary relation, thus reading aPb as "alternative a is strictly preferred to b". Let $\mathcal{L}(\mathcal{X})$ denote the set of all linear orders on $\mathcal{X}$. Aggregation procedures in this framework are functions $F : \mathcal{L}(\mathcal{X})^{\mathcal{N}} \to \mathcal{L}(X)$ and are called *social welfare functions* (SWFs).

4.1 Translation

Let us now consider the following setting for binary aggregation: define a set of issues $\mathcal{I}_{\mathcal{X}}$ as the set of all pairs (a, b) in $\mathcal{X}$. The domain $\mathcal{D}_{\mathcal{X}}$ of aggregation is therefore $\{0, 1\}^{|\mathcal{X}|^2}$. In this setting a binary ballot corresponds to a binary relation P over $\mathcal{X}$: $B_{(a,b)} = 1$ iff a is in relation to b (aPb). Given this representation, we

can associate with every SWF for $\mathcal{X}$ and $\mathcal{N}$ an aggregation procedure on a subdomain of $\mathcal{D}_{\mathcal{X}}^{\mathcal{N}}$.

Using the propositional language $\mathcal{L}_{PS}$, we can express properties of binary ballots in $\mathcal{D}_{\mathcal{X}}$. In this case the language consists of $|\mathcal{X}|^2$ propositional symbols, which we shall call p_{ab} for every issue (a, b). The properties of linear orders can be enforced on binary ballots using the following set of integrity constraints, which we shall call $\mathrm{IC}_<$:[3]

Completeness and antisymmetry:
$p_{ab} \leftrightarrow \neg p_{ba}$ for $a \neq b \in \mathcal{X}$ $\quad$ $\neg p_{aa}$ for all $a \in \mathcal{X}$

Transitivity: $p_{ab} \wedge p_{bc} \rightarrow p_{ac}$ for $a, b, c \in \mathcal{X}$ pairwise distinct

Note that the size of this set of integrity constraints is polynomial in the number of alternatives in $\mathcal{X}$.

It is now straightforward to see that every SWF corresponds to an aggregation procedure that is collectively rational wrt. $\mathrm{IC}_<$, and *vice versa*. Moreover, if the SWF satisfies the unanimity axiom of preference aggregation (Gaertner 2006), then the associated binary aggregation procedure satisfies unanimity as defined in Section 2.3. The same is true for the axioms of anonymity, independence, and monotonicity (but note that for the two axioms of neutrality the correspondence is not straightforward).

4.2 Condorcet Paradox and Impossibilities

The translation presented above enables us to express the famous Condorcet Paradox in terms of Definition 2.1. Let $\mathcal{X} = \{a, b, c\}$ and let $\mathcal{N}$ contain three individuals. Consider the following profile $\mathbf{B}$, where we have omitted the values of the reflexive issues *aa* (always 0 by $\mathrm{IC}_<$), and specified the value of only one of *ab* and *ba* (the other can be obtained by taking the opposite of the value of the first):

	ab	*bc*	*ac*
Agent 1	1	1	1
Agent 2	0	1	0
Agent 3	1	0	0
Majority	1	1	0

[3] We will use the notation IC both for a single integrity constraint and for a set of formulas—in the latter case considering as the actual constraint the conjunction of all the formulas in IC.

Clearly, every individual ballot satisfies $\mathrm{IC}_<$, but the outcome obtained by taking majorities violates one formula, namely $p_{ab} \wedge p_{bc} \rightarrow p_{ac}$. Therefore, $(F_{\mathrm{maj}}, \mathbf{B}, \mathrm{IC}_<)$ is a paradox by Definition 2.1, where F_{maj} is the majority rule.

Now, by a syntactic analysis of the transitivity constraints introduced before, we can observe that they are in fact equivalent to just two positive clauses: The first one, $p_{ab} \vee p_{bc} \vee p_{ca}$, rules out the cycle $a<b<c<a$, and the second one, $p_{ba} \vee p_{cb} \vee p_{ac}$, rules out the opposite cycle $c<b<a<c$. That is, these constraints correspond exactly to the two Condorcet cycles that can be created from three alternatives.

We will now show how characterisation results of CR procedures for specific propositional languages, such as those given in Section 3, can be used to prove impossibility theorems in preference aggregation, similar to Arrow's Theorem (Arrow 1963). Call an SWF *imposed* if for some pair of distinct alternatives a and b we have that a is always collectively preferred to b in every profile.

Proposition 11. *If $|\mathcal{X}| \geqslant 3$ and $|\mathcal{N}| \geqslant 2$, then any anonymous, independent and monotonic SWF for $\mathcal{X}$ and $\mathcal{N}$ is imposed.*

Proof. In the first part of Section 4 we have seen that every anonymous, independent and monotonic SWF corresponds to a binary aggregation procedure that is collectively rational for $\mathrm{IC}_<$ and that satisfies A, I and M. By Proposition 1, every A, I, M aggregation procedure is a quota rule. We will now prove that, if a quota rule is collectively rational for $\mathrm{IC}_<$, then it is imposed, i.e., at least one of the quotas q_{ab} is equal to 0.

Suppose, for the sake of contradiction, that every quota $q_{ab} > 0$. As remarked before, for any three alternatives $a, b, c \in \mathcal{X}$ the integrity constraints corresponding to transitivity are $p_{ba} \vee p_{ca} \vee p_{bc}$ and $p_{ab} \vee p_{ac} \vee p_{cb}$. These are positive clauses of size 3; thus, by Proposition 10 we obtain:

$$q_{ab} + q_{bc} + q_{ca} < n + 3$$
$$q_{ba} + q_{cb} + q_{ac} < n + 3$$

Furthermore, it is easy to see that the $\mathrm{IC}_<$ for completeness and antisymmetry force the quotas to satisfy the following: $q_{ab}+q_{ba}=n+1$, $q_{bc}+q_{cb}=n+1$, and $q_{ac}+q_{ca}=n+1$.

Now, adding the two inequalities we obtain that $\sum_{a,b\in\mathcal{X}} q_{ab} < 2n + 6$ and adding the three equalities we obtain $\sum_{a,b\in\mathcal{X}} q_{ab} = 3n + 3$. The two constraints together admit a solution only if $n < 3$. Thus, it remains to analyse the case of 2 individuals; but it is easy to see that our constraints do not admit a solution in positive integers for $n = 2$. This shows that there must be a quota $q_{ab} = 0$ for certain distinct a and b as soon as $n \geqslant 2$; hence, the SWF is imposed. □

Arrow's Theorem states that every SWF satisfying U and I is dictatorial, and, although intuitively stronger, it does not imply Proposition 11. The importance of our result lies in the structure of its proof: most proofs of Arrow's Theorem and similar results concentrate on so-called "decisive coalitions". Here instead we point out a clash between axiomatic requirements and the syntactic shape of integrity constraints.

5 Judgment Aggregation

In this section we review the framework of *judgment aggregation* (List and Puppe 2009), and we provide a characterisation of judgment aggregation procedures as collectively rational procedures wrt. a particular set of integrity constraints.

Judgement aggregation (JA) considers problems where a finite set of individuals $\mathcal{N}$ has to generate a collective judgment over a set of interconnected propositional formulas Φ. Formally, we call *agenda* a finite nonempty set Φ of propositional formulas, not containing any doubly-negated formulas, that is closed under complementation (i.e, $\alpha \in \Phi$ whenever $\neg\alpha \in \Phi$, and $\neg\alpha \in \Phi$ for every positive $\alpha \in \Phi$). Each individual in $\mathcal{N}$ expresses a *judgment set* $J \subseteq \Phi$, as the set of those formulas in the agenda that she judges to be true. Every individual judgment set J is assumed to be *complete* (i.e., for each $\alpha \in \Phi$ either α or its complement are in J) and *consistent* (i.e., there exists an assignment that makes all formulas in J true). If we denote by $\mathcal{J}(\Phi)$ the set of all complete and consistent subsets of Φ, we can define a *JA procedure* for Φ and $\mathcal{N}$ as a function $F : \mathcal{J}(\Phi)^{\mathcal{N}} \to 2^{\Phi}$. A JA procedure is called *complete* (resp. *consistent*) if the judgment set it returns is complete (resp. consistent) on every profile.

5.1 Translation

Let us now consider the following binary aggregation framework. Let the set of issues $\mathcal{I}_\Phi$ be equal to the set of formulas in Φ. The domain $\mathcal{D}_\Phi$ of aggregation is therefore $\{0,1\}^{|\Phi|}$. In this setting, a binary ballot corresponds to a judgment set: $B_\alpha = 1$ iff $\alpha \in J$. Given this representation, we can associate with every JA procedure for Φ and $\mathcal{N}$ a binary aggregation procedure on a subdomain of $\mathcal{D}_\Phi^{\mathcal{N}}$. Note that this translation is different from the one given by Dokow and Holzman (2010), which deals with models of judgment sets (rather than judgment sets) as input of the aggregation.

As before, we now define a set of integrity constraints for $\mathcal{D}_\Phi$ to enforce the properties of consistency and completeness. The propositional language in this

case consists of $|\Phi|$ propositional symbols p_α, one for every $\alpha \in \Phi$. Recall that a *minimally inconsistent set* (mi-set) of propositional formulas is an inconsistent set each proper subset of which is consistent. Let IC_Φ be the following set of integrity constraints:

Completeness: $p_\alpha \vee p_{\neg\alpha}$ for all $\alpha \in \Phi$

Consistency: $\neg(\bigwedge_{\alpha\in S} p_\alpha)$ for every mi-set $S \subseteq \Phi$

Note that the size of IC_Φ might be exponential in the size of the agenda. This is in agreement with considerations of computational complexity: Since checking the consistency of a judgment set is an NP-hard problem, while model checking on binary ballots is in P, the translation from JA to binary aggregation must contain an exponential step.

The same kind of correspondence we have shown for SWFs holds between complete and consistent JA procedures and binary aggregation procedures that are collectively rational with respect to IC_Φ. We also obtain a perfect correspondence between the axioms, as every unanimous (resp. anonymous, independent, neutral, monotonic) JA procedure corresponds to a unanimous (resp. anonymous, independent, issue-neutral, monotonic) binary aggregation procedure.

5.2 Doctrinal Paradox and Agenda Properties

The paradox of JA was first studied in the literature discussing legal doctrines and then formalised in JA under the name of Doctrinal Paradox (List and Puppe 2009). Let Φ be the agenda $\{\alpha, \beta, \alpha \wedge \beta\}$[4] and let $\mathbf{B}$ be the following profile:

	α	β	$\alpha \wedge \beta$
Agent 1	1	1	1
Agent 2	0	1	0
Agent 3	1	0	0
Majority	1	1	0

Every individual ballot satisfies IC_Φ, while the outcome contradicts the constraint $\neg(p_\alpha \wedge p_\beta \wedge p_{\neg(\alpha\wedge\beta)}) \in \mathrm{IC}_\Phi$. Hence, $(F_{\mathrm{maj}}, \mathbf{B}, \mathrm{IC}_\Phi)$ constitutes a paradox by Definition 2.1.

The notion of *safety of the agenda* introduced in previous work (Endriss et al. 2010) is related to our definition of paradox. An agenda Φ is *safe* wrt. a class of

[4] We omit negated formulas; for any $J \in \mathcal{J}(\Phi)$ their acceptance can be inferred from the acceptance of the positive counterparts.

JA procedures if any procedure in the class will return consistent outcomes for any profile over Φ. Several characterisation results have been proved that links agenda properties ensuring safety and classes of procedures defined axiomatically. As we shall see next, the translation of the JA framework into binary aggregation enables us to obtain a syntactic analogue of these properties. To simplify presentation, we shall assume that agendas do not include tautologies (or contradictions).

We say that an agenda Φ satisfies the *syntactic simplified median property* (SSMP) if every mi-subset of Φ is of the form $\{\alpha, \neg\alpha\}$. This corresponds to IC_Φ being equivalent to the conjunction of $p_\alpha \leftrightarrow \neg p_{\neg\alpha}$ for all positive $\alpha \in \Phi$. A weaker condition is the *simplified median property* (SMP), which holds if every mi-subset of Φ is of the form $\{\alpha, \neg\beta\}$ for α logically equivalent to β. Equivalences between formulas are expressed using bi-implications; thus, the SMP corresponds to adding to the previous set of constraints a set of positive bi-implications $p_\alpha \leftrightarrow p_\beta$ for any equivalent α and β in Φ. These considerations enable us to give a new proof for and strengthen a result that was proved in previous work (Endriss et al. 2010, Theorem 8). Call a procedure *complement-free* if the outcome never includes two formulas that are (syntactic) complements, for any profile in $\mathcal{J}(\Phi)^N$.

Proposition 12. *An agenda Φ is safe for the class of complete, complement-free, and neutral JA procedures if and only if Φ satisfies the SMP.*

Proof. By translating JA into binary aggregation we have that Φ is safe wrt. complete, complement-free and neutral JA procedures if and only if IC_Φ does not generate a paradox with any issue-neutral procedure. It is easy to see that complete and complement-free procedures are characterised by procedures that are CR wrt. to constraints of the form $p_\alpha \leftrightarrow \neg p_{\neg\alpha}$. Therefore, we can concentrate on the remaining condition. We know by Proposition 3 that an issue-neutral procedure is collectively rational for IC_Φ iff IC_Φ is expressible in $\mathcal{L}_\leftrightarrow$, and using our earlier syntactic characterisation we conclude that this is the case iff Φ satisfies the SMP. □

The statement of Proposition 12 drops the axiom of anonymity, which was assumed in the previous statement of the theorem, and it does not require a representation result for its proof.

5.3 Another Characterisation Result: the Majority Rule

We will now glance back at the lifting results of Section 3, obtaining a characterisation of the set of integrity constraints that are lifted by the majority rule by exploiting the link between binary aggregation and JA. A result proved by Nehring and Puppe (2007) in the framework of JA shows that the majority rule is consistent if and only if the agenda Φ satisfies the *median property*, i.e., if there exists no mi-subset of Φ of size greater than 2. Binary aggregation problems with integrity constraints can be viewed as JA over atomic agendas: a ballot over issues $i_1, \ldots, i_m$ can be viewed as a complete judgment set over a set of propositional symbols $p_1, \ldots, p_m$, the consistency of a judgment set being defined as consistency *with respect to* the constraint IC. Ballots are assignments that may satisfy or falsify IC. Therefore, a mi-subset of the agenda corresponds to a *minimally falsifying partial assignment* (mifap-assignment) for IC: an assignment to some of the propositional variables that cannot be extended to a satisfying assignment, although each of its proper subsets can. Therefore, we obtain the following characterisation:

Lemma 4. *The majority rule is CR wrt. to* IC *if and only if there is no mifap-assignment for* IC *of size greater than 2.*

Let us now prove a crucial lemma about mifap-assignments. Associate with each mifap-assignment ρ a conjunction $C_\rho = \ell_1 \wedge \cdots \wedge \ell_k$, where $\ell_i{=}p_i$ if $\rho(p_i){=}1$ and $\ell_i{=}\neg p_i$ if $\rho(p_i){=}0$, for all propositional symbols p_i on which ρ is defined.

Lemma 5. *Every non-tautological formula φ is equivalent to $(\bigwedge_\rho \neg C_\rho)$ with ρ ranging over all mifap-assignments of φ.*

Proof. Let A be a total assignment for φ. Suppose $A \nvDash \varphi$, i.e., A is a falsifying assignment for φ. Since φ is not a tautology there exists at least one such A. By sequentially deleting propositional symbols from the domain of A we find a mifap-assignment ρ_A included in A. Hence, A falsifies the conjunct associated with ρ_A, and thus the whole formula $(\bigwedge_\rho \neg C_\rho)$.

Assume now $A \models \varphi$ but $A \nvDash (\bigwedge_\rho \neg C_\rho)$. Then there is a ρ such that $A \models C_\rho$. This implies $\rho{\subseteq}A$, and since ρ is a mifap-assignment for φ this contradicts the assumption $A \models \varphi$. □

Proposition 13. *The majority rule is CR wrt.* IC *if and only if* IC *is expressible as a conjunction of clauses of size $\leqslant 2$.*

Proof. The proof for one direction can be found in previous work (Grandi and Endriss 2010, Proposition 18): the majority rule is CR wrt. conjunctions of 2-clauses. The other direction is entailed by the two lemmas above: Suppose that the majority rule is CR wrt. IC, then, by Lemma 4, IC does not have any mifap-assignment of size > 2. Therefore, by Lemma 5, we can construct a conjunction of 2-clauses that is equivalent to IC, as every conjunct C_ρ in the statement of Lemma 5 has size $\leqslant 2$. The case of IC being a tautology is straightforward, as every tautology is equivalent to a 2-clause, namely $p \vee \neg p$. □

6 Conclusions and Future Work

We introduced a simple propositional language to express individual rationality constraints in the framework of binary aggregation, and we defined an aggregation procedure to be collectively rational if the collective outcome satisfies a certain constraint whenever all individuals do. We proved several results to characterise, for various subsets of the language, a set of axioms that guarantees the collective rationality of a procedure for all constraints in this subset, and we have outlined an approach for how to apply these results in more complex situations. We have explored the potential of the framework of binary aggregation with integrity constraints as a general framework for the analysis of collective choice problems, by showing how two of the main frameworks of Social Choice Theory, preference and judgment aggregation, can be embedded into binary aggregation by defining suitable integrity constraints. We were able to give new and simpler proofs of theoretical results in both frameworks, and to characterise seemingly unrelated paradoxes as instances of the same general definition.

This work can be extended in a number of ways. The first step towards a generalisation to the case of full (rather than boolean) combinatorial domains (Lang 2007; 2004) is a study of the case of *voting for committees*, where the domain is a product space of domains D of equal cardinality. By defining a language from propositional symbols $\{p_{x_j=a} \mid a \in D, j \in \mathcal{I}\}$ it is possible to generate integrity constraints to model various voting procedures, such as approval voting, and prove preliminary results linking axioms with syntactic requirements on additional integrity constraints. Another direction is to allow for sequential aggregation procedures: by analysing the integrity constraints we might be able to devise a meaningful agenda for the decision process. Finally, by using more powerful languages to express rationality assumptions we can move towards more complex logical models of artificial agents.

Acknowledgements The authors would like to thank the anonymous referees of AAAI-2010 and IJCAI-2011 for their precious comments, as well as the audience of SCW-2010, the COMSOC seminar and the LIRa seminar at ILLC, and the Doctoral School on Computational Social Choice in Estoril, where different versions of this paper have been presented.

References

K. J. Arrow. *Social Choice and Individual Values*. John Wiley & Sons, 2nd edition, 1963.

Y. Chevaleyre, U. Endriss, J. Lang, and N. Maudet. Preference handling in combinatorial domains: From AI to social choice. *AI Magazine*, 29(4):37–46, 2008.

F. Dietrich and C. List. Judgment aggregation by quota rules: Majority voting generalized. *Journal of Theoretical Politics*, 19(4):391–424, 2007.

E. Dokow and R. Holzman. Aggregation of binary evaluations. *Journal of Economic Theory*, 145(2):495–511, 2010.

U. Endriss, U. Grandi, and D. Porello. Complexity of judgment aggregation: Safety of the agenda. In *Proceedings of the 9th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS-2010)*, 2010.

W. Gaertner. *A Primer in Social Choice Theory*. Oxford University Press, 2006.

U. Grandi and U. Endriss. Lifting rationality assumptions in binary aggregation. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence (AAAI-2010)*, 2010.

U. Grandi and U. Endriss. Binary aggregation with integrity constraints. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI-2011)*, 2011. To appear.

J. Lang. Logical preference representation and combinatorial vote. *Annals of Mathematics and Artificial Intelligence*, 42(1–3):37–71, 2004.

J. Lang. Vote and aggregation in combinatorial domains with structured preferences. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-2007)*, 2007.

C. List and C. Puppe. Judgment aggregation: A survey. In *Handbook of Rational and Social Choice*. Oxford University Press, 2009.

K. O. May. A set of independent necessary and sufficient conditions for simple majority decision. *Econometrica*, 20(4):680–684, 1952.

K. Nehring and C. Puppe. The structure of strategy-proof social choice. Part I: General characterization and possibility results on median spaces. *Journal of Economic Theory*, 135(1):269–305, 2007.

R. B. Wilson. On the theory of aggregation. *Journal of Economic Theory*, 10(1): 89–99, 1975.

A Dynamic Epistemic Logic Approach to Modeling *Obligationes*

Sara L. Uckelman

Institute for Logic, Language & Computation
Universiteit van Amsterdam
S.L.Uckelman@uva.nl

Abstract

The game-like nature of the medieval theory of *obligationes* is well-recognized. In an *obligatio*, two agents, the Opponent and the Respondent, engage in a turn-based dialogue, where the Respondent's actions are governed by certain rules, and the goal of the dialogue is establishing the consistency of a proposition. Given the central importance of the Opponent, the Respondent, and the rules governing the turn-taking and actions of both players, it seems natural to look to game-based structures in logic to provide a general framework for modeling different types of *obligationes*. However, few attempts have been made to provide an explicit specification of the game(s) which are involved in different types of obligational disputations. Such formalizations which have been previously proposed are specific to the particular obligational framework being investigated, and have little in terms of explanatory value concerning the foundation of truth and knowledge within the disputation. In this paper we present an alternative framework, which we call the *reductive framework*, which can be used to model many different types of *obligationes* within a single system.

1 Introduction

The game-like nature of the medieval theory of *obligationes* has been well-recognized (de Rijk 1975, Dutilh Novaes 2007, Hamblin 1970). In an *obligatio*, two agents, the Opponent and the Respondent, engage in a turn-based dialogue, where the Respondent's actions are governed by certain rules, and the goal of the dialogue is establishing the consistency of a proposition. Given the central importance of the Opponent, the Respondent, and the rules governing the turn-taking and actions of both players, it seems natural to look to game-based structures in logic to provide a general framework for modeling different types of *obligationes*. In particular, the apparent similarity between *obligationes* and Lorenzen's dialogical semantics is immediate. However, few attempts have been made to provide an explicit specification of the game(s) which are involved in different types of obligational disputations. Such logical models which have been previously proposed are usually designed specifically to model the particular obligational framework being investigated, and have little in terms of explanatory value concerning the foundation of truth and knowledge within the disputation. This is because they are in general *additive* models, that is, they start from a small base (usually a single formula) and *add* to that base as the disputation proceeds.

In this paper we present an alternative which can be used to model many different types of *obligationes* within a single framework and which is *reductive* in nature, that is, the starting model of a disputation will be large, and the actions of the players will successively reduce the model. This allows us to model knowledge and logical consequence explicitly. This approach, based on a multi-agent variant of Dynamic Epistemic Logic (DEL) (van Ditmarsch et al. 2007), is also abstract enough to be able to model multiple types of *obligationes* within the same general framework, merely by selecting different classes of models satisfying different properties. The result is a precise mathematical model which will allow us to give formal proofs of properties of *obligationes* and thus to shed light on some of the puzzling aspects of medieval obligational theories which arise from their presentation in semi-formal natural language rather than in symbolic, mathematical language.

The plan of the paper is as follows. In the next section, we give a general introduction to the medieval theory of *obligationes*, illustrated by the specific example of Walter Burley's theory of *positio*. In §3, we consider some aspects of *obligationes* which are not obviously present in dialogical games of Lorenzen's logic or in game-theoretic semantics. In §4 we introduce the additive framework for modeling *obligationes*, the *status quo*. We point out some of its deficiencies,

and introduce our alternative in §5. In §6 we apply our framework to various types of *obligationes*, and in §7 we conclude.

2 The medieval theory of *obligationes*

An obligational disputation, or *obligatio*, is a dialogue between two agents, the Opponent and the Respondent, where the Opponent puts forward a sequence of propositions, and the Respondent is obligated (hence the name) to follow certain rules in his responses to the Opponent's propositions. More precisely, the Opponent puts forward an initial statement, called the *positum*, which the Respondent can either accept or refuse to accept. If he accepts, the *obligatio* begins. If he does not, no *obligatio* begins. If the *obligatio* begins, the Opponent puts forward propositions and the Respondent has three ways that he can respond: He can grant or concede (*concedere*) the proposition, he can deny (*negare*) the proposition, or he can doubt (*dubitare*) it, that is, remain agnostic. (Some authors, such as William of Ockham (of Ockham 1974) and the anonymous author of the *Obligationes Parisienses* (de Rijk 1975), mention a fourth option, which is to 'draw distinctions' (*distinguere*), that is, to clarify an ambiguity on the part of the Opponent, but for ease of modeling we will ignore this action in the present paper.) The *obligatio* continues until the Opponent calls "*Cedat tempus*" ("Time's up"), whereupon the responses of the Respondent are analysed with respect to the rules the Respondent was supposed to follow, to determine whether the Respondent has responded well or badly.

The earliest texts on *obligationes* date from the early 13th century de Rijk (1974; 1975; 1976), and while their roots are clearly grounded in Aristotle's discussion of dialectical exchanges in the *Topics* VIII, 4 (159a15–24) and in the *Prior Analytics* I, 13 (32a18–20) (cf. Yrjönsuuri 1994, §II.A), the systematic development of the theory of *obligationes* over the course of the 13th and 14th centuries tends to show little adherence to the Aristotelian definitions. While the specific details vary from author to author, a number of distinct types of *obligationes* discussed by multiple authors can be identified. The six most common are: *positio*, *depositio*, *dubitatio*, *sit verum* or *rei veritatis*, *institutio*, and *petitio*. Of these six, *positio* is universally the most widely studied, both by medieval and modern authors.[1] (For a more detailed introduction to *obligationes*, including a

[1]The additive frameworks for modeling *obligationes* that we study below focus specifically on *positio*. There is little previous work specifically discussing *dubitatio*. Because of the interesting unique properties that *dubitatio* has, we must ensure that our framework is able to model this type of obligational disputation. In §6 we briefly discuss the details of modeling *dubitatio*, which details

discussion of their purpose and their role in medieval academics, see Yrjönsuuri (1994).)

2.1 Walter Burley's *obligationes*

Walter Burley's treatise *De obligationibus*, written around 1302, when he was a master of arts at the University of Oxford, gives a standard treatment of *positio*. The text of this treatise is edited in Burley (1963) and a partial translation of the text, including the section on *positio* in its entirety, is found in Burley (1988). Burley defines the general goal of an *obligatio* as follows:

> The opponent's job is to use language in a way that makes the respondent grant impossible things that he need not grant because of the *positum*. The respondent's job, on the other hand, is to maintain the *positum* in such a way that any impossibility seems to follow not because of him but rather because of the *positum* (Burley 1988, p. 370).[2]

Thus, it is clear that in an *obligatio*, the goal is not to demonstrate the logical truth or validity of the initial proposition, but instead to maintain a level of consistency throughout the entire disputation. Burley gives three general rules that must always be adhered to in any *obligatio* (Burley 1988, p. 375):

General Rule 1: Everything following from an *obligatum* must be granted (where '*obligatum*' is interpreted as what has been granted or what must necessarily be granted).

General Rule 2: Everything incompatible with the *obligatum* must be denied.

General Rule 3: One must reply to what is irrelevant in accordance with its own quality.

Definition 2.1. (Relevance). A proposition is *irrelevant* or *impertinent* if neither it nor its negation follows from the set of propositions which have already been conceded (which includes the negations of propositions which have been denied).

---

are fully worked out in Uckelman (2011).

[2] *Opus opponentis est sic inducere orationem ut faciat respondentem concedere impossibilia quae propter positum non sunt necessaria concedere. Opus autem respondentis est sic sustinere positum ut propter ipsum non videatur aliquod impossibile sequi, sed magis propter positum* (Burley 1963, p. 34).

In *positio*, the primary obligation of the Respondent is to grant, that is, to hold as true, the *positum*, the initial statement put forward by the Opponent. If the Respondent accepts the *positum* and the *obligatio* begins, the additional rules that he must follow are:

Rule 1: Everything that is posited and put forward in the form of the positum during the time of the positio must be granted (Burley 1988, p. 379).[3]

Rule 2: Everything that follows from the positum must be granted. Everything that follows from the positum either together with an already granted proposition (or propositions), or together with the opposite of a proposition (or the opposites of propositions) already correctly denied and known to be such, must be granted (Burley 1988, p. 381).[4]

Rule 3: Everything incompatible with the positum must be denied. Likewise, everything incompatible with the positum together with an already granted proposition (or propositions), or together with the opposite of a proposition (or the opposites of propositions) already correctly denied and known to be such, must be denied (Burley 1988, p. 381).[5]

Rules 1, 2, and 3 are just precisifications of General Rules 1 and 2. In Rule 1, 'in the form of' should be understood syntactically; if the *positum* is 'Marcus is Roman', then the Respondent isn't obliged to accept 'Tullius is Roman' unless it is explicit (either through common knowledge or through previous concessions) that Marcus is Tullius. Note that even though Burley's rules include reference to the epistemic states of the Respondent, in practice he generally ignores these epistemic clauses (Yrjönsuuri 1994, p. 53). Later authors who include these clauses do take them seriously; we discuss this in §6.4 below.

2.2 Example *positiones*

We illustrate Burley's rules for *positio* with two examples, one very simple, the other illustrating the interesting higher-order nature of some obligational disputations.

[3] *Omne positum, sub forma positi propositum, in tempore positionis, est concedendum* (Burley 1963, p. 46).

[4] *Omne sequens ex posito est concedendum. Omne sequens ex posito cum concesso vel concessis, vel cum opposito bene negati vel oppositis bene negatorum, scitum esse tale, est concedendum* (Burley 1963, p. 48).

[5] *Omne repugnans posito est negandum. Similiter omne repugnans posito cum concesso vel concessis, vel opposito bene negati vel oppositis bene negatorum, scitum esse tale, est negandum* (Burley 1963, p. 48).

	Opponent	**Respondent**
1	I posit φ.	I admit it.
2	$\neg\varphi \vee \psi$.	I concede it.
3	ψ	I concede it.

Figure 1: A simple *positio*.

	Opponent	**Respondent**
1	I posit φ or φ must be granted.	I admit it.
2	φ must be granted.	I deny it.
3	φ follows from the positum and the opposite of something correctly denied	I grant it.
4	φ must be granted.	??

Figure 2: A more interesting *positio*.

For the first example, given in Figure 1, suppose that φ does not imply $\neg\psi$ and φ is known to be false. In the first round, the Opponent puts forward a contingent (but false) proposition φ; the Respondent grants it in accord with Rule 1. In the second round, either φ implies ψ, then the sentence is relevant and follows from the set of propositions conceded so far along with the negations of propositions denied to this point; or it doesn't, in which case it is irrelevant and true (since φ is false). In both cases, the Respondent is required to concede; the first case falls under Rule 2, and the second under General Rule 3, the rule for irrelevant propositions. In the third round, the Respondent likewise must concede because ψ follows from the Respondent's new commitment set. This disputation shows how, given a *positum* which is false, but not inconsistent, the Opponent can force the Respondent to concede any other consistent proposition.

A more interesting example is given in Figure 2, and involves statements about the obligational rules themselves. Let φ be the proposition 'you are in Rome' (spoken by the Opponent to the Respondent). The *positum* is a disjunction between a simple proposition and the assertion that that proposition must be granted. Because the disjunction is not a logical contradiction (in particular the first disjunct is possible, though it is in fact false), the Respondent is correct in accepting the *positum*. The second disjunct is irrelevant, as it is not a logical

consequence of the *positum*, and furthermore it is false: Since φ is false, and φ is also irrelevant, the Respondent is not under any obligation to accept φ. Thus it is false that φ must be granted, so he correctly denied the second proposition. The third proposition expresses a logical necessity, about the validity of disjunctive syllogism, and so is accepted. But now it is unclear how the Respondent should respond to the re-assertion that φ must be granted. On the one hand, this proposition has been put forward before, and was denied, and so it should continue to be denied. On the other hand, once the third proposition has been granted, by Rule 2, φ must be granted. So superficially it appears that the Respondent is obliged to both accept and deny this final statement. Burley's resolution to the problem is to argue that (3) is not only not necessary, but it is repugnant, since it is inconsistent with the opposite of (2). Since it is repugnant, the Respondent should have in fact denied, and thus (4) can also be denied without contradiction (Yrjönsuuri 1994, pp. 152–155).

3 Unique issues in modeling *obligationes*

There are two dominant paradigms in semantic games in logic: the dialogical games of Lorenzen (Lorenzen 1955) and the game-theoretic semantics (GTS) of Hintikka (Hintikka and Sandu 1997). In both these paradigms, there are three components to the game: two players (usually called the Opponent and the Proponent), the set of possible moves that each player can (legally) make, and a set of strategies (i.e., sequences of moves), of which a subset are identified as winning strategy. The set of possible moves of the players can be seen as a definition of the semantics of the logical connectives (cf. Rahman and Keiff 2005, p. 365, or Pietarinen 2006, p. 325). Two of these components are also present in *obligationes*: two players (called the Opponent and the Respondent), and the set of possible moves that each player can (legally) make. The definition of a 'winning strategy' in an *obligatio* is a problem we discuss in §3.2. Yet despite the strong game-like appearance of obligational disputations, and their apparent similarity to dialogues, it is not immediately clear whether it is possible to model *obligationes* in either of the two paradigms. We sketch briefly some of the features of *obligationes* which are not obviously present in standard game/dialogue approaches.

3.1 Closed- vs. open-world models

In dialogical logic, the Proponent puts forward a proposition which he will try to defend. The two players proceed in sequential turns, with both the Opponent and the Proponent having a set of attacks and a set of defenses that they may use against actions in the previous round(s). In the presentation of Lorenzen's dialogical logic by Rahman and Keiff (2005), dialogues of infinite length are allowed (cf. [§2.3]). However, it is possible to specify termination conditions, under which the dialogue could continue, but in which if it did, it would never generate any state other than a state already reached. With the introduction of a formal definition of strict repetition and rules which disallow strict repetition in a dialogue, it is possible to define the validity of a proposition in terms of the existence of a winning strategy for the Proponent **P**:

Definition 3.1. (Validity). In a given dialogical system the proposition expressed by the formula stating the thesis is said to be valid iff **P** has a (formal) winning strategy for it, i.e., **P** can in accordance with the appropriate rules succeed in defending the thesis against all possible allowed criticism by **O** (Rahman and Keiff 2005, p. 369).

The first point to note is that this type of approach—defining validity in terms of winning strategies that depend on the disallowance of strict repetitions—only works in what is called a *closed-world* situation. In these dialogical games, the Proponent is not allowed to assert any atom which the Opponent has not already asserted, and hence one feature of a successful strategy for the Proponent will be to force the Opponent into asserting as many atoms as possible. The Opponent wants to assert as few atoms as possible, and in particular, he never has any incentive to assert any atom which does not occur in the formula which is under discussion. In contrast, *obligationes* are *open-world* dialogues, which is to say that it can be advantageous to the Opponent to introduce formulas into the disputation which involve proposition letters not present in the *positum*. In fact, not only is this possible, it is nearly ubiquitous, for without the addition of new matter into the disputation, it may be difficult for the Opponent to trap the Respondent into responding badly.

3.2 Winning conditions

Because *obligationes* are open-world dialogues rather than closed-world dialogues, it is possible that the games last an infinite amount of time, without being repetitious in the way defined above. However, in practice, the Opponent

will always call "*Cedat tempus*" after a finite amount of time has elapsed. This makes it tricky to define the winning conditions for the Respondent. Clearly, the Opponent has won if, when he calls "*Cedat tempus*", the Respondent has conceded a contradiction, or has both conceded and denied the same proposition. (We make this more explicit below). However, just because the Respondent has not conceded a contradiction after a finite number of steps is no guarantee that *positum* is consistent. It is always possible that the Opponent has not yet introduced new atoms which would cause the Respondent's downfall.

3.3 Asymmetry of roles

Pietarinen argues that a multiplicity of types of games, including dialogical logic and game-theoretic semantics, can be modeled by the activities of seeking and finding, or of showing what one sees.[6] A consequence of such an approach is that:

> Unlike examinations, the activities of seeking and finding, or showing what one sees, are not asymmetric, although they may well be either cooperative or non-cooperative in nature (Pietarinen 2006, p. 337).

Obligationes are essentially asymmetric, in that the rules governing the behavior of the Opponent and the Respondent are disjoint.[7] The Respondent never asserts any statement of his own devising, he only ever responds to propositions put forward by the Opponent. Or, to look at it from a different perspective, the Opponent never *asserts* any statement, he only proffers them, and while the Respondent's action of 'concede' or 'grant' can be interpreted as an assertion on the part of the Respondent, the Respondent will never assert any statement not first proffered by the Opponent. When viewed in this way, *obligationes* look much less like dialogues than they may initially, since there is no set of commitments, either positive or negative, that the Opponent is bound to (cf. Karunatillake et al. 2009, Maudet and Chaib-Draa 2002).

[6] Cf. Angelelli (1970, p. 802), where *obligationes* are classified under the "question" method of disputation, rather than the "argument" method.

[7] In fact, in most texts, rules for the Opponent are not given at all. One exception is the early text *Tractatus Emmeranus* (de Rijk 1974), which gives some rules (better thought of as guidelines, or strategic advice) to the Opponent.

3.4 Goal of the games

Most importantly, whatever type of game *obligationes* turn out to be, they are not going to be semantic games in the same way that GTS and dialogical logic are. In both GTS and dialogical logic, the goal of the Proponent is to prove the logical validity of the formula under dispute. The validity or the truth of the formula is not known in advance, it is only known after the game has terminated. Additionally, as we noted at the beginning of this section, the attack and defense rules for Lorenzen dialogues are used to define the meaning of the logical connectives. Neither of these is the case in *obligationes*. Many treatises on *obligationes* point out that the truth-value of the *positum* should be known in advance (at least by the Opponent).[8] An *obligatio* that starts with a true *positum* will never be of interest; instead, the Opponent should try to pick a *positum* which is known to be false, but not impossible. (If it is impossible, it will be fairly easy for the Opponent to force the Respondent into conceding a contradiction, in which case the Respondent should not have accepted that proposition as the *positum* in the first case. It is in this way that *obligationes* can be seen as proof of *possible*, as opposed to *actual*, truth.) Thus at least certain types of *obligationes*, such as *positio*, are best understood as games of consistency, rather than of validity. Similarly, the rules for the connectives are not given via the rules for the Respondent, but instead the Respondent must know in advance the rules governing logical consequence, in order to be able to follow the rules (cf. Angelelli 1970, p. 813). The fact that many semantic properties of the propositions involved need to be known in advance is the main argument against the adequacy of the additive framework for modeling *obligationes*, and in favor of the reductive framework that we introduce below.

4 The additive framework

Given Burley's statement about the goal of an *obligatio*, it is natural to model Burley-style obligations as a type of *consistency-maintenance* game. Such a model is proposed in Dutilh Novaes (2007, §3.3). Formally, the model is a structure $\mathfrak{M}^{O} = \langle K_c, \Phi, \Gamma, R(\varphi)\rangle$ where K_c is the set of common knowledge among the participants of the disputation (expressed as a set of propositions); Φ is a sequence of propositions, each element of which is denoted φ_n, which keeps track of the assertions of the Opponent; Γ is a sequence of propositions, which

[8] Thus, it is clear that *obligationes* should not be classified as "proof games", which demonstrate logical truth or validity, as Pietarinen does (Pietarinen 2006, p. 319).

keeps track of the responses of the Respondent; and $R(\varphi)$ is a function from φ to the values 1, 0, and ?, indicating that the Respondent has granted, denied, or doubted the proposition φ. The rules of the obligation can thus be seen as the constraints on the type of function that R can be. Dutilh Novaes gives the following formalization of Burley's rules:

Definition 4.1. The *logical rules* of Burley's *positio* are as follows.[9] If φ_0 = the *positum*, then

$$R(\varphi_0) = \begin{cases} 0 & \text{iff } \varphi_0 \vdash \bot \\ 1 & \text{iff } \varphi_0 \nvdash \bot \end{cases}$$

For $\varphi_n, n > 0$:

$$R(\varphi_n) = \begin{cases} 0 & \text{iff } \Gamma_{n-1} \vdash \neg\varphi_n, \text{ or} \\ & \Gamma_{n-1} \nvdash \varphi_n, \Gamma_{n-1} \nvdash \neg\varphi_n \text{ and } K_C \vDash \neg\varphi_n \\ 1 & \text{iff } \Gamma_{n-1} \vdash \varphi_n, \text{ or} \\ & \Gamma_{n-1} \nvdash \varphi_n, \Gamma_{n-1} \nvdash \neg\varphi_n \text{ and } K_C \vDash \varphi_n \\ ? & \text{iff } \Gamma_{n-1} \nvdash \varphi_n \end{cases}$$

Definition 4.2. (Formation of Γ_n). The set Γ_n is formed inductively. $\Gamma_{0-1} = \emptyset$, and:

$$\Gamma_n = \begin{cases} \Gamma_{n-1} \cup \{\varphi_n\} & \text{if } R(\varphi_n) = 1 \\ \Gamma_{n-1} \cup \{\neg\varphi_n\} & \text{if } R(\varphi_n) = 0 \\ \Gamma_{n-1} & \text{if } R(\varphi_n) = ? \end{cases}$$

It is clear from this exposition that the model starts with a small base, namely $\Gamma_0 = \{\varphi_0\}$ = the *positum*, and grows as the disputation proceeds. We call logical models of this type *additive*, since as the number of rounds of the disputation increases, so does the size of the model. These additive structures can be easily adapted to satisfy different sets of rules, such as those of Richard Swyneshed (c.1330), who defines pertinence solely in terms of the *positum*, and not the set of propositions already conceded or denied, and Ralph Strode (second half of the 14th C), who introduces epistemic clauses into the rules. Such adaptations can be found in Dutilh Novaes (2007, §3.4) and Dutilh Novaes (2007, §3.5), respectively.

When *obligationes* are modeled in an additive framework, it is hard to classify them as either dialogical games of validity or semantic games of model-building, since application of the rules depends crucially on having a semantic

[9] We correct the notation of Dutilh Novaes (2007, p. 156) by using $\vdash$ ('derives') rather than $\Vdash$ ('forces'), since it is clearly the model-theoretic, not the set-theoretic, notion that is being used.

model already in hand. The additive framework presuppose a significant amount of background information which is never specified: the semantic model(s) in which truth of propositions (particularly the *positum* and irrelevant propositions) and the Respondent's knowledge of both individual propositions as well as consequent relations is to be evaluated, and the syntactic rules governing $\vdash$. The main drawback of the additive framework is that this information is taken for granted. For example, the set of common knowledge K_C is not defined in any explicit fashion, and there is nothing which grounds the knowledge of the participants:

> K_C is the common state of knowledge of those present at the disputation complemented by the *casus*... [It] is an incomplete model, in the sense that some propositions do not receive a truth-value: for some propositions, it is not known whether they are true or false... although it may be known that they are true-or-false. So, the state of common knowledge is a state of imperfect information (Dutilh Novaes 2007, p. 155).

Additionally, since the nature of the proof system being used in the definition of $R(\varphi_n)$ is never specified, the additive framework is essentially incomplete; it is impossible to implement the logical model without making the proof-system explicit (Dutilh Novaes 2007, p. 169). The reductive framework that we present in the next section addresses these drawbacks.

5 The reductive framework

In the reductive approach to modeling *obligationes*, the underlying logic is multi-agent Dynamic Epistemic Logic (DEL, van Ditmarsch et al. (2007)). We first introduce static Epistemic Logic (EL) and then add dynamics to form the basis of our reductive framework. (Since in what follows we will only be working with the multi-agent versions, we will no longer specify this explicitly.) Before applying this framework to various types of *obligationes* in §6, we discuss motivations for our choice of logic, and how it differs in important respects from better known fragments of DEL.

5.1 The logic

A multi-agent epistemic logic is an extension of propositional logic with a family of modal operators K_a for $a \in \mathcal{A}$. We are interested in a particular extension of

standard epistemic logic, namely, *epistemic logic with common knowledge*, which has a further family of operators C_G, for $G \subseteq \mathcal{A}$. For a set Φ_0 of propositional letters and set $\mathcal{A}$ of agents, the set Φ_{EL} of wffs of EL is defined by:

$$\varphi \quad := \quad p \in \Phi_0 \mid \neg\varphi \mid \varphi \vee \varphi \mid K_a\varphi : a \in \mathcal{A} \mid C_G\varphi : G \subseteq \mathcal{A}$$

$K_a\varphi$ is read 'agent a knows that φ'. $C_G\varphi$ is read 'it is common knowledge amongst the group of agents G that φ'. We will use C_G to represent the knowledge of the two agents at the beginning of the disputation.

Epistemic logic is interpreted with Kripke models.

Definition 5.1. (Epistemic models). A structure $\mathfrak{M} = \langle W, w^*, \{\sim_a : a \in \mathcal{A}\}, V\rangle$ is an *epistemic model* if

- W is a set, with $w^* \in W$ a designated point (representing the actual world).
- $\{\sim_a : a \in \mathcal{A}\}$ is a family of equivalence relations on W, one for each member of $\mathcal{A}$. The relation $w \sim_a w'$ is interpreted as 'w and w' are epistemically equivalent for agent a'; that is, a cannot distinguish between w and w'. $\sim_G : G \subseteq \mathcal{A}$ is defined as the transitive closure of $\bigcup_{a\in G}\{\sim_a\}$.
- $V : \Phi_0 \to 2^W$ is a valuation function associating atomic propositions with subsets of W. For $p \in \Phi_0$, if $w \in V(p)$, we say that 'p is true at w' and write $\mathfrak{M}, w \vDash p$.

We designate the class of epistemic models by $\mathcal{M}$.

Definition 5.2. (Semantics). The semantics for the propositional connectives are as expected. We give just the semantics for the epistemic operators.

$$\begin{array}{lll} \mathfrak{M}, w \vDash K_a\varphi & \text{iff} & \forall w'(\langle w, w'\rangle \in \sim_a \text{ implies } \mathfrak{M}, w' \vDash \varphi) \\ \mathfrak{M}, w \vDash C_G\varphi & \text{iff} & \forall w'(\langle w, w'\rangle \in \sim_G \text{ implies } \mathfrak{M}, w' \vDash \varphi) \end{array}$$

Epistemic models cover the knowledge of the agents; to model their actions, we add dynamics, via Propositional Dynamic Logic (PDL). PDL is an extension of propositional logic by a family of modal operators $[\alpha]$ for $\alpha \in \Pi$, a set of programmes (or more generally, a set of actions or events). The language of PDL is two-sorted, with a set Φ_0 of atoms and a set Π_0 of atomic actions. We do not need the full expressivity of PDL to model *obligationes*, so we introduce only the fragment we require (for the full version, see Harel et al. (2002)). We let $\Pi_0 = \emptyset$, and the sets Φ_{Ob} and Π_{Ob} of complex well-formed formulas and programmes are defined by mutual induction:

$$\begin{array}{lll} \varphi & := & \varphi \in \Phi_{\text{EL}} \mid [\alpha]\varphi : \alpha \in \Pi_{\text{Ob}} \\ \alpha & := & \varphi? : \varphi \in \Phi_{\text{EL}} \end{array}$$

The programme $\varphi?$ is to be interpreted as a test operator, which tests for the truth of φ. Note that the only programmes that we allow are testing of formulas which do not themselves contain any programmes. The semantics for the new $[\alpha]$ operator are given in terms of model reduction. Let $\mathfrak{M} \restriction \varphi = \langle W^{\mathfrak{M},\varphi}, \{\sim_a^{\mathfrak{M},\varphi} : a \in \mathcal{A}\}, V^{\mathfrak{M},\varphi}\rangle$, where $W^{\mathfrak{M},\varphi} := \{w \in W : \mathfrak{M}, w \vDash \varphi\}$, and the relations and valuation functions are just restrictions of the originals. For a set of ordered propositions Γ_n, let $\mathfrak{M} \restriction \Gamma_n = \mathfrak{M} \restriction \gamma_0 \restriction \cdots \restriction \gamma_n$, that is, $\mathfrak{M} \restriction \Gamma_n$ is the result of the sequential restriction of $\mathfrak{M}$ by the elements of Γ_n. Then:

Definition 5.3 (Semantics)**.**

$$\mathfrak{M}, w \vDash [\varphi?]\psi \quad \text{iff} \quad \forall v \in \mathfrak{M} \restriction \varphi, v \vDash \psi$$

The dual test programme $\langle\varphi?\rangle$ has the following semantics:

$$\mathfrak{M}, w \vDash \langle\varphi?\rangle\psi \quad \text{iff} \quad \exists v \in \mathfrak{M} \restriction \varphi, v \vDash \psi$$

These truth conditions are reductive, in that an announcement (i.e., an action by the Respondent) results in the reduction of the original model to a smaller model.

5.2 Discussion

The dynamic aspect of reasoning in obligational disputations is perhaps the most important feature of *obligationes*. This dynamic shift is described by Ekenberg when he says:

> It is the granting of the disjunction that is the critical step. If A is the sentence 'you are in Rome' and B is the sentence 'you are a bishop', what happens is that the semantic background connected to the component A in the disjunction shifts from being something related to the knowledge of the respondent to being strictly determined by the *positum*. As the *positum* is contrary to fact, this will inevitably lead to a difference in truth-value before and after the granting of the disjunction (Ekenberg 2002, p. 30).

It is interesting, then, to ask how the dynamics involved in *obligationes* differ from modern dynamical systems. The fragment of DEL introduced in the previous section is similar to, though not the same as, a familiar fragment much discussed in recent literature, Public Announcement Logic (PAL, (van

Ditmarsch et al. 2007, ch. 4)). In PAL, "public announcements" $[\varphi]\psi$ are modeled via the test operator, with the added condition that the announcement (the formula being tested for) must be true at the world of evaluation:

$$\mathfrak{M}, w \vDash [\varphi]\psi \quad \text{iff} \quad \mathfrak{M}, w \vDash \varphi \quad \text{and} \quad \mathfrak{M} \upharpoonright \varphi, w \vDash \psi$$

That is, the actual world (the world of evaluation) will always remain in the model after the model reduction operation has been performed. In contrast, in our framework, if the Opponent has picked his *positum* correctly, then the actual world will be removed from the model with the first action of the Respondent (this will be made more precise in the next section).

At this point it is worth commenting on our choice of logical framework, since DEL is not the only dynamic approach to modeling knowledge change out there. One alternative approach is update semantics (US), developed in Veltman (1996). In US, the "meaning of a sentence is an operation on information states" (Veltman 1996, p. 221), and the update operation is one that superficially looks similar to the model reduction operation by which we defined the truth conditions for the test programme. The propositions by which information states are updated can be thought of as observations given by nature—with the Opponent in an *obligatio* playing the role of "nature".[10] But there are a number of methodological reasons not to pursue this route.

First, reducing the Opponent to the role of "nature" hides the multi-agent aspect of *obligationes*. When building a formal model of a historical theory, one cannot overlook or ignore the component features of that theory. In treatises *de obligationibus*, the two-player character of the disputation is always emphasized. Any modern formal model which hopes to do justice to the medieval theory must take into account the features which the medieval texts emphasize. While the interaction between the two players in an *obligatio* is mostly one-sided (more *re*action than *inter*action), Opponent's role in the *obligatio* is not negligible: It is his decision when to end the disputation, and if Respondent is trapped into contradiction it is as much a result of Opponent's cleverness as it is Respondent's inability to follow the rules correctly. Hence, if we want our formal model to be faithful to the original theories, these facts should not be obfuscated in the model that we construct.

Second, the US framework does not allow for the addition of incorrect information, which prevents us from modeling the Respondent's acceptance of a false *positum*. Veltman distinguishes between *propositional updates*, which are additive, and *tests*, which are not (Veltman 1996, p. 225). A propositional

[10] We thank an anonymous commentator on an earlier version of this paper for this observation.

update is like a public announcement in that they both must be true at the actual world in order for them to be successful. Tests, on the other hand, are non-monotonic:

> [T]he outcome of this test can be positive at first and negative later. In the minimal state you have to accept 'It might be raining', but as soon as you learn that it is not raining 'It might be raining' has to be rejected (Veltman 1996, p. 229)

This type of nonmonotonicity is important in US because one intended application of US is modeling default reason, which is generally defeasible. However, this is not the type of reasoning that goes on in *obligationes*. A general characteristic shared by almost all types of *obligationes*, *positio* and otherwise, is that if the Respondent follows the rules correctly, the only way he will give different answers to the same proposition at different rounds of the disputation is if he first doubts the proposition, and then at a later stage concedes or denies it. Thus, the obligational systems are monotonic in so far as the only change in Respondent's response is moving from doubt concerning a specific proposition to certainty (that is, concession or denial).

Third, knowledge in US is not necessarily veridical; it could be that what an agent believes she knows is in fact false (Veltman 1996, fn. 260). In *obligationes*, however, knowledge is generally considered nondefeasible and unrevisable: It is never lost, only gained (though in some contexts, such as *dubitatio*, it can be tabled temporarily).

For these reasons, we find a DEL-based approach preferable.

6 Applications

In this section, we show how the reductive framework can model various aspects of *obligationes*. Throughout, our set of agents is $A = \{\mathsf{Opp}, \mathsf{Res}\}$ (for Opponent and Respondent, respectively).

Definition 6.1. (Actions of Res). Let φ_n be a proposition put forward by Opp. The possible actions of Res (designated Act) are:

concede:	$[\varphi_n]\top$
deny:	$[\neg\varphi_n]\top$
doubt:	$[\top]\top$

These actions are essentially testing for consistency. The last clause in this definition is equivalent to saying "I don't know"; $[\top]\top$ will always be valid, in any epistemic model.

Definition 6.2. (*Obligatio*). An *obligatio* is a quadruple $O = \langle \Theta, R, \Gamma, \Gamma^R \rangle$ where

- Θ is a sequence of propositions, such that $\theta_0 \in \Theta$ is the *obligatum* and $\theta_n \in \Theta$ is the proposition put forward by Opp at round n.
- $R : \Theta \times \mathbb{N} \to \text{Act}$ is a function determining Res's correct response to each element of Θ. We write $R(\theta_n)$ for $R(\theta, n)$ to simplify notation.
- Γ is a sequence of actions, formed by Res's actual responses to each element of Θ.
- Γ^R is a sequence of actions, formed by the *correct* response of Res to each element in Θ, as given by R.

We will often abuse notation and identify Γ and Γ^R, which are sequences of tests, with the sequences of tested formulas (that is, if $\Gamma_1 = \langle [\theta_0?]\top, [\neg\theta_1?]\top \rangle$, we identify Γ_1 with $\langle \theta_0, \neg\theta_1 \rangle$). The actions of Res successively reduce the model; it is this dynamic process which introduces the difficulty for the Respondent, in terms of keeping track of his responses, real-world facts, and the inferential relationships between propositions.

In *obligationes* where the goal is consistency maintenance, there are two different notions of winning that we may be interested in, a global notion and a local notion.

Definition 6.3. (Winning, global). Opp *wins* if there is some n such that $\mathfrak{M} \upharpoonright \Gamma_n = \langle \emptyset, \{\sim_a^{\mathfrak{M},\Gamma_n} : a \in A\}, V^{\mathfrak{M},\Gamma_n} \rangle$. Res wins otherwise.

Opp wins if he succeeds in making Res answer in a contradictory fashion. The only time W will be empty is when Γ_n is inconsistent. Once the model has been reduced to the worlds which satisfy φ, reducing it to the worlds which satisfy $\neg\varphi$ will result in the empty set.

We may also be interested in a slightly weaker notion of winning, which can be applied to an individual *obligatio*, which will always be finite:

Definition 6.4. (Winning, local). If Opp calls "*Cedat tempus*" after round n, then Opp *wins* if

$$\mathfrak{M} \upharpoonright \Gamma_n = \langle \emptyset, \{\sim_a^{\mathfrak{M},\Gamma_n} : a \in A\}, V^{\mathfrak{M},\Gamma_n} \rangle$$

and Res wins otherwise.

6.1 An example: Burley's *positio*

Different types of *obligatio* and different author's versions of *obligatio* are modeled by putting constraints on R and these constraints combined with Θ will generate Γ^R. We give a simple example, that of Burley-style *positio* (cf. §2): $O^{\text{Bur}} = \langle \Theta, R^{\text{Bur}}, \Gamma, \Gamma^{R^{\text{Bur}}} \rangle$, where Θ can be any sequence of propositions, since there are no constraints on Opp's behavior, and Γ is any sequence of actions, since Res can respond in any fashion that he wishes to Opp's proposals (but it is only if he follows the set of rules that he actually plays the game correctly). The rules he must follow in Burley's *positio* are defined as follows:

Definition 6.5. For a model $\mathfrak{M}$ and formula $\theta_0 \in \Theta$ = the *positum*:

$$R^{\text{Bur}}(\theta_0) = \begin{cases} \textbf{concede} & \text{iff } \mathfrak{M}, w \vDash \langle \theta_0 \rangle \top \\ \textbf{deny} & \text{iff } \mathfrak{M}, w \vDash [\theta_0] \bot \end{cases}$$

For $\theta_n \in \Theta, n > 0$:

If $\mathfrak{M} \restriction \Gamma_{n-1} \vDash \theta_n$:	$R^{\text{Bur}}(\theta_n) =$ **concede**
If $\mathfrak{M} \restriction \Gamma_{n-1} \vDash \neg\theta_n$:	$R^{\text{Bur}}(\theta_n) =$ **deny**
Otherwise:	
If $\mathfrak{M}, w^* \vDash K_{\mathsf{Res}}\theta_n$:	$R^{\text{Bur}}(\theta_n) =$ **concede**
If $\mathfrak{M}, w^* \vDash K_{\mathsf{Res}}\neg\theta_n$:	$R^{\text{Bur}}(\theta_n) =$ **deny**
If $\mathfrak{M}, w^* \vDash \neg(K_{\mathsf{Res}}\theta \vee K_{\mathsf{Res}}\neg\theta_n)$:	$R^{\text{Bur}}(\theta_n) =$ **doubt**

Since Burley's rules are deterministic, the sequence $\Gamma^{R^{\text{Bur}}}$ is uniquely defined.

Definition 6.6. The sequence $\Gamma_n^{R^{\text{Bur}}}$ of *rule-following responses of* Res is formed as follows:

$$\begin{aligned} \Gamma_0^{R^{\text{Bur}}} &= \langle R^{\text{Bur}}(\theta_0) \rangle \\ \Gamma_n^{R^{\text{Bur}}} &= \langle \gamma_0, \ldots, \gamma_{n-1}, R^{\text{Bur}}(\theta_n) \rangle \end{aligned}$$

The rule-following sequences are relevant when we consider, in future work, questions of the computational complexity of *obligationes*.

6.2 Complex *posita*

A natural variant on Burley's *positio* is considered by Boethius de Dacia (third quarter of the 13th C) in his *Questiones super librum Topicorum* (de Dacia et al. 1976). It is:

> based on the opponent positing every thesis which he wants to posit, and the respondent must grant them, whether they are probable or improbable, whether necessary or impossible, as far as it does not happen that they are incompossible—incompossibility being the only cause why the respondent has to deny the opponent any of those, which he wants to posit (Yrjönsuuri 1994, p. 32).

Formally, instead of putting forward a single *positum* φ_0, Opp instead puts forward a set Ψ of propositions. Let $\varphi_0 = \bigwedge \Psi$, and then Boethius's rules for R and Γ^R follow Burley's (cf. Yrjönsuuri 1994, p. 59)).

6.3 Explicit common knowledge

In an arbitrary epistemic model $\mathfrak{M}$, the set of propositions which are common knowledge amongst a group of agents is not explicitly specified. In an *obligatio*, the set of common knowledge, against which the truth of irrelevant propositions is evaluated, is likewise often left implicit. In some cases, before the *obligatio* begins, a *casus* is introduced. A *casus* is a hypothesis about how the world is, or extra information about how the *positum* should be analyzed (Yrjönsuuri 1993). In the first sense, the *casus* can be understood as a set of literals expressing the *explicit common knowledge* at the start of the game, so the *casus* can be implemented by a restriction on V. We therefore implement the restriction on V as follows:

Definition 6.7. (*Casus*). Let $\mathrm{Lit}(\Phi_0)$ be the set of literals of Φ_0, and $C \subseteq \mathrm{Lit}_{\Phi_0}$ be the *casus*. Then $\mathfrak{M}$ *models the casus* if there is a partition $P = P_1 \cup P_2$ of W with P_1 containing w^*, such that if $w \sim_{\mathsf{Res}} w^*$, then $w \in P$, if $v \sim_{\mathsf{Opp}} w^*$, then $v \in P_1$, and for all $w, v \in P_1$, $w \sim_{\mathsf{Res}} v$ and $w \sim_{\mathsf{Opp}} v$; and for every positive literal $p \in C$ and every $w \in P_1$, $w \in V(p)$, and for every negative literal $\neg q \in C$ and every $w \in P_1$, $w \notin V(q)$.

Note that we allow that the *positum* contradicts information in the *casus*. In principle, it is not required that the *casus* be consistent, but for the purposes of modeling we disregard ones which are not, because if the *casus* is not consistent, then Res should not accept the *positum*, since Opp could easily force him into conceding a contradiction, and so no *obligatio* would begin. Thus, if we are interested in modeling actual *obligationes*, we do not need to have a provision for ones which have inconsistent *casus*.

Lemma 1. *If* $\mathfrak{M}$ *models a* casus C*, then for every* $\varphi \in C$, $\mathfrak{M} \vDash C_{\{\mathsf{Opp},\mathsf{Res}\}}\varphi$.

Proof. Straightforward. □

6.4 Logical omniscience

Related to the question of explicit common knowledge is the issue of logical omniscience. Epistemic models satisfy the property of *logical omniscience*, that is, for any $a \in A$, $\varphi, \psi \in \Phi_{\mathrm{Ob}}$, and arbitrary model $\mathfrak{M}$, the following holds:

$$\mathfrak{M} \vDash K_a\varphi \wedge K_a(\varphi \to \psi) \to K_a\psi$$

Logical omniscience is problematic when modeling resource-bounded agents, such as agents participating in certain types of *obligatio*. As we said earlier, in simple variants of *positio* such as Burley's the presence of logical omniscience is not problematic, since the epistemic clauses in the rules are not taken seriously. However, some later authors do take them seriously. For example, Richard Brinkley (3rd quarter of the 14th C) gives the following rules for pertinent propositions (Brinkley et al. 1995, p. 15):

> Everything following from the *positum* proposed during the time of its *positio* and known to be such must be conceded.
>
> Everything incompatible with the *positum* during the time of the *positio* and known to be such must be denied.

In order for these rules to be relevantly different from Burley's, it has to be the case that there are some valid inferences that Res doesn't recognize, that is, he cannot be logically omniscient. Fagin et al. (1995) introduce ways of constructing agents who are not logically omniscient. One way is through the addition of 'awareness' functions for each agent to the model.

Definition 6.8. (Awareness). E_a is an *awareness function for a* if $E_a \colon W \to 2^{\Phi_{\mathrm{Ob}}}$. Intuitively, $E_a(w)$ is the set of propositions that a is aware of at w (cf. Fagin et al. 1995, §9.5).

Definition 6.9. A structure $\mathfrak{M}_E = \langle \mathfrak{M}, \{E_a \colon a \in \mathcal{A}\}\rangle$ is an *epistemic model with explicit awareness* if $\mathfrak{M}$ is an epistemic announcement model and each E_a is an awareness function. We denote the class of these models by $\mathcal{M}_E$.

For an arbitrary $\mathfrak{M}_E \in \mathcal{M}_E$, implicit knowledge is defined as knowledge in $\mathfrak{M}$ and explicit knowledge as implicit knowledge which the agent is also aware of. The introduction of awareness functions for Opp and Res allows us to model Brinkley-style rules, with epistemic clauses for the players.

6.5 Drawing distinctions

In §2 we mentioned that some medieval authors allow Res to draw distinctions, that is, to clarify ambiguity introduced by Opp. One way to understand this drawing of distinctions is as role-switching. When Opp puts forward a proposition φ_n that Res finds ambiguous, Res can treat φ_n as if Opp has in fact put forward a (finite) *set* Δ_{φ_n} of propositions (the set of possible ways to understand φ_n). Then, the current *positio* is paused while Res takes on the role of Opp and vice versa. Res then selects $\psi \in \Delta_{\varphi_n}$, and proffers it. He continues to select a unique ψ until Opp concedes one of them, in which case the roles reverse, the agents return to the original *positio*, and Res responds to the output of the procedure. Since Δ_{φ_n} is finite, this procedure will terminate at some point. If Opp's intended reading of φ_n is not in Δ_{φ_n}, then he will never concede any proposition put forward by Res, and then the proper response for Res is to doubt φ_n.

The analogy of role-switching is, however, not completely accurate. First, there is no initial statement that Opp must concede before the new game begins. Second, there are no rules governing Opp's responses to the propositions Res proffers him; hence, there is no obligation. Still, this analogy provides a convenient way of understanding, formally, what is going on when Res is drawing distinctions.

An alternative way of understanding drawing distinctions is found in post-medieval treatises on disputations, which follow in the footsteps of *obligationes*. In this tradition, "*distinguo*" can be analysed as a meta-level response which divides the game into a set of two or more arguments, one for each sense of the proposition φ_n, which Res and Opp then play simultaneously (Angelelli 1970, pp. 808–809). Analysed in this fashion, *distinguo* is similar to having a disjunctive *posita*. Disjunctive *posita* are discussed by Burley, who considers them to be a species of indeterminate *positio* (Burley 1963, p. 73–74). If the game opens by Opp saying "Either I posit φ or I posit ψ", then "the respondent is, on Burley's interpretation, left in doubt about what is his positum, and therefore he has to answer with doubt both to φ and to ψ, if they are false, since he does not know which of them is his positum" (Yrjönsuuri 1994, p. 59). Alternatively, Res could play two games at the same time, one with φ as his *positum* and one with ψ.

6.6 *Dubitatio*

In the variants of *positio* that we've discussed, the obligations of Res all affect his propositional attitudes towards propositions (even though those propositions themselves may involve higher-order attitudes, such as knowledge). In *dubitatio*, the obligation of the Res is relevantly different: His primary obligation is to doubt the initial statement (the *dubitatum*), and to ensure that all his successive answers preserve this doubt; that is, he should either doubt or deny any statement implying the *dubitatum*, as well as deny or concede any statement implied by the *dubitatum*. *Dubitatio* is often considered to be a "trivial variation on positing [*positio*]" Spade (2003), by both modern and medieval authors.[11] However, as is argued in (Uckelman et al. 2011), this is too dismissive of a view. Just as in *positio*, where the *positum* must be false in order for the disputation to be interesting, in *dubitatio*, the *dubitatum* should be something whose truth value is in fact known by Res; if the *dubitatum* is already doubtful for Res, then there is nothing particularly tricky about the disputation. Given an arbitrary *positio* and model, if the *positum* is consistent, it will always be possible to find a world in a model where the *positum* is true (and if Opp has chosen correctly, it will be a world other than the actual world). Thus it is straightforward in *positio* to say precisely how Res should act if he wants to follow the rules, because Res's moves are uniquely determined by the rules.

This is not the case with *dubitatio*. Since *dubitatio* deals not with the object-level of truth or falsity of propositions, but instead with the meta-level of knowledge of propositions, *dubitatio* cannot be handled in the same straightforward fashion. In an arbitrary model, for any given proposition which is known by Res at the actual world, there may be multiple ways of changing the model to result in one where that proposition is no longer known. However, there is very little discussion of this issue in the literature on standard dynamic epistemic logic, since knowledge as we defined it in §5 is generally considered to be "hard" knowledge, indefeasible and irrevocable (cf. Baltag and Smets (2009, p. 126) and van Ditmarsch (2005, pp. 230, 235)). Because it is assumed that this type of knowledge cannot be lost once it has been gained, little consideration is given to how to model updates which move an agent from a situation where φ is known to one where φ is not known. In other work (Uckelman 2011) we have developed a plausible type of model update that can be used to model this type of situation; such an update procedure is of interest beyond just modeling

[11] For example, Paul of Venice agrees with Spade's assessment of *dubitatio*; he says that "every *dubitatio* or *petitio* is a *positio*, as was shown at the beginning" (of Venice and Ashworth (ed. and trans.) 1988, p. 327).

dubitatio since it can be used to model agents who wish to hide their knowledge from other agents, and must interact and reason in such a way as to not divulge that they know certain propositions.

7 Conclusions

Hamblin, one of the first modern writers to recognize the game-like and dialogue-like nature of *obligationes*, said in 1970:

> The Game of Obligation. . . has been replaced by nothing else and, although it was never developed at a very high theoretical level, there is every reason to take it seriously and try to learn from it something relevant to modern times (Hamblin 1970, p. 125).

In this paper we have surveyed previous work on formal modeling of *obligationes*, which is additive in nature, and noted the drawbacks of the additive approach. To ameliorate these drawbacks, we have introduced the reductive framework based on multi-agent Dynamic Epistemic Logic, and showed how this framework can be used to model certain aspects of different variants of *positio* in a unified fashion. There is much work still to be done to provide explicit specifications for different types of *obligationes*, including variants other than *positio*, and we are excited that this work will help shed new light on our understanding of this fascinating medieval genre.

Acknowledgements This research was funded by the project "Dialogical Foundations of Semantics" (DiFoS) in the ESF EuroCoRes programme LogICCC (LogICCC-FP004; DN 231-80-002; CN 2008/08314/GW). Earlier versions of this paper were presented in the Logic and Interactive Rationality seminar, Amsterdam, March 2010, and as a tutorial at the Dialogues and Games workshop, Lille, February 2010, and the author would like to thank audience members for useful feedback. The author is also grateful for the comments of anonymous referees on an earlier version of this paper.

References

I. Angelelli. Techniques of disputation in the history of logic. *Journal of Philosophy*, 67(20):800–815, 1970.

A. Baltag and S. Smets. Learning by questions and answers: From belief-revision cycles to doxastic fixed points. In H. Ono, M. Kanazawa, and R. de Queiroz, editors, *Logic, Language, Information, and Computation: Proceedings of the 16th International Workshop, WoLLIC 2009, Tokyo, Japan, June 21–24, 2009*, LNCS 5514, pages 124–139. Springer, 2009.

R. Brinkley, P. V. Spade (ed.), and G. A. Wilson (ed.). *Richard Brinkley's Obligationes: A late fourteenth century treatise on the logic of disputation*. Beiträge zur Geschichte der Philosophie und Theologie des Mittelalters. Aschendorff, 1995.

W. Burley. Tractatus de obligationibus. In R. Green, O.F.M., editor, *An Introduction to the Logical Treatise 'De Obligationibus'*, volume 2. Université Catholique, 1963.

W. Burley. Obligations (selections). In N. Kretzmann and E. Stump, editors, *The Cambridge Translations of Medieval Philosophical Texts*, volume 1: logic and the philosophy of language, pages 369–412. Cambridge University Press, 1988.

B. de Dacia, N. Green-Pedersen (ed.), and J. Pinborg (ed.). *Questiones Super Librum Topicorum*, volume 6 of *Corpus Philosophorum Danicorum Medii Aevi*. Danish Soc. of Language and Literature, 1976.

L. M. de Rijk. Some thirteenth century tracts on the game of obligation. *Vivarium*, 12(2):94–123, 1974.

L. M. de Rijk. Some thirteenth century tracts on the game of obligation II. *Vivarium*, 13(1):22–54, 1975.

L. M. de Rijk. Some thirteenth century tracts on the game of obligation III. *Vivarium*, 14(1):26–49, 1976.

C. Dutilh Novaes. *Formalizing Medieval Logical Theories: Suppositio, Consequentiae and Obligationes*, volume 7 of *Logic, Epistemology, and the Unity of Science*. Springer, 2007.

T. Ekenberg. Order in obligational disputations. *Disputatio*, 5:23–39, 2002.

R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. *Reasoning About Knowledge*. MIT Press, 1995.

C. L. Hamblin. *Fallacies*. Methuen, 1970.

D. Harel, D. Kozen, and J. Tiuryn. Dynamic logic. In D. Gabbay and F. Guenther, editors, *Handbook of Philosophical Logic*, volume 4. Kluwer, 2 edition, 2002.

J. Hintikka and G. Sandu. Game-theoretical semantics. In J. van Benthem, J. ter Meulen, and A. ter Meulen, editors, *Handbook of Logic and Language*, pages 361–410. MIT Press, 1997.

N. C. Karunatillake, N. R. Jennings, I. Rahwan, and P. McBurney. Dialogue games that agents play within a society. *Artificial Intelligence*, 173:935–981, 2009.

P. Lorenzen. *Einführung in die operative Logik und Mathematik*. Springer, 1955.

N. Maudet and B. Chaib-Draa. Commitment-based and dialogue-game-based protocols: New trends in agent communication languages. *Knowledge Engineering Review*, 17(2):157–179, 2002.

W. of Ockham. *Opera Philosophica*, volume I. Franciscan Institute, 1974.

P. of Venice and E. J. Ashworth (ed. and trans.). *Logica Magna Part II, Fascicule 8: Tractatus de Obligationibus*. Oxford University Press, 1988.

A.-V. Pietarinen. *Signs of Logic: Peircean Themes on the Philosophy of Language, Games, and Communication*. Springer, 2006.

S. Rahman and L. Keiff. On how to be a dialogician. In D. Vanderken, editor, *Logic, Thought, and Action*, volume 2 of *Logic, Epistemology, and the Unity of Science*, pages 359–408. Springer, 2005.

P. V. Spade. Medieval theories of obligationes. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2003 edition, 2003. `http://plato.stanford.edu/archives/fall2003/entries/obligationes/`.

S. L. Uckelman. Deceit and indefeasible knowledge: The case of *Dubitatio*. In submission, 2011.

S. L. Uckelman, J. Maat, and K. Rybalko. The art of doubting in *Obligationes Parisienses*. In C. Kann, B. Löwe, C. Rode, and S. L. Uckelman, editors, *Modern Views of Medieval Logic*, Recherches de Théologie et Philosophie Médiévales—Bibliotheca. Peeters, 2011.

H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*, volume 337 of *Synthese Library Series*. Springer, 2007.

H. P. van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese*, (147):229–275, 2005.

F. Veltman. Defaults in update semantics. *Journal of Philosophical Logic*, 25: 221–261, 1996.

M. Yrjönsuuri. The role of casus in some fourteenth century treatises on sophismata and obligations. In K. Jacobi, editor, *Argumentationstheorie. Scholastische Forschungen zu den logischen und semantischen Regeln korrekten Folgerns*, pages 301–321. Brill, 1993.

M. Yrjönsuuri. *Obligationes: 14th Century Logic of Disputational Duties*, volume 55 of *Acta Philosophica Fennica*. Societatis Philosophia Fennica, 1994.

Reasoning with Protocols under Imperfect Information

Eric Pacuit and Sunil Simon

Tilburg Center for Logic and Philosophy of Science (TiLPS), Centrum Wiskunde en Inforamtica (CWI)
e.j.pacuit@uvt.nl,s.e.simon@cwi.nl

Abstract

Various combinations of temporal logics, epistemic and doxastic logics, and action logics have been used to reason about (groups of) agents in social situations. A key issue that has emerged is how best to represent and reason about the underlying *protocol* that governs the agents' interactions in a particular social situation. In this paper, we propose a PDL-style logic for reasoning about protocols under imperfect information.

Our paper touches on a number of issues surrounding the relationship between an agent's abilities, available choices and information in an interactive situation. The main question we address is under what circumstances can the agent commit to a protocol or plan, and what can she achieve by doing so?

1 Introduction and motivation

There is a growing literature using different (multi-)modal logics to reason about communities of agents engaged in some form of social interaction. In particular, various combinations of temporal logics, epistemic and doxastic logics, action logics and preference logics have been studied in this context[1].

[1] A complete survey of these "logics of rational agency" is outside the scope of this paper. The interested reader can consult van Benthem (2010), van der Hoek and Wooldridge (2003a), Meyer

A key issue that has emerged is how best to represent and reason about the underlying *protocol* that governs the agents' interactions in a particular social situation.

Intuitively, a *protocol* describes what the agents "can" or "cannot" do (say, observe) in a social interactive situation. This leads to *substantive* assumptions about the formal model, such as which actions (observations, messages, utterances) are available (permitted) at any given moment. These assumptions can be roughly categorized according to the different uses of "can":

(1) To describe physical, temporal or historical possibilities: A typical example is the assumption an agent *cannot* receive a message unless another agent sent it earlier. Such assumptions limit the options available to the agents at any given moment.

(2) To describe the agents' abilities, or skills: The options available to an agent at any given moment are defined not only by what is "physically possible," but also by the agent's *capacity* to perform various actions. For example, "Ann *can* throw a bulls-eye" typically means that Ann has the ability to (repeatedly) throw a bulls-eye.

(3) To describe compliance to some type of norm: The social or conversational[2] norms at play in the interactive situation being modeled (i.e., the "rules of the game") impose further constraints on the options available to each agent. For example, common conversational rules include: "Do not blurt everything out at the beginning"; "Do not repeat yourself"; "Let others speak in turn"; and "Be honest." Imposing such rules *restricts* the legitimate sequences of possible statements.

So, a protocol encodes not only which options are *feasible*, but also what is *permissible* for the agents to do or say. Of course, an interesting and important component of a logical analysis of rational agents is to disambiguate these different meanings of "can" (cf. Horty 2001, van der Hoek et al. 1998, Elgesem 1997, Governatori and Rotolo 2005, Cross 1986). In this paper, we take a more abstract perspective in which a protocol simply identifies a subtree from the "grand stage" of all possible sequences of events that could take place in an interactive situation.

A number of authors have forcefully argued that the underlying protocol is an important component of any analysis of (social) interactive situations and

and Veltman (2007) for a discussion and for references to the relevant literature.

[2]See (Parikh and Ramanujam 2003, Section 6) for a discussion of Gricean norms in this context.

should be explicitly represented in a formal model (cf. Fagin et al. 1995, van Benthem et al. 2009, Parikh and Ramanujam 2003, Hoshi 2009, Wang 2010). Indeed, much of the work over the past 20 years using epistemic logic to reason about distributed algorithms has provided interesting case studies highlighting the interplay between protocol analysis and epistemic reasoning (an important example here is the seminal paper by Halpern and Moses (1990) on the "generals problem").

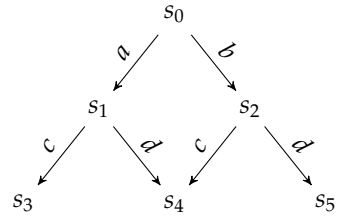
The central question of this paper is what do the agents "know" about the underlying protocol, and how is this reflected in the logic used to reason about social interactions? A typical assumption is that there is a fixed, global protocol that all the agents have (explicitly or implicitly) agreed to follow (and this is commonly known). This is the assumption in the *epistemic temporal logics*, as discussed by Parikh and Ramanujam (2003), Halpern and Moses (1990), van Benthem et al. (2009), among many others (Fagin et al. 1995, van Benthem 2010, are textbook presentations of this literature). These logical systems use linear or branching time models with added epistemic structure induced by the agents' different capacities for observing events. The models provide a "grand stage" where histories of some social interaction unfold constrained by an underlying *protocol*. Thus, the protocol is represented *extensionally* in the models as a set of histories (sequences of events)[3]. From the point of view of the logical systems that have been developed to reason about these structures (e.g., as discussed in Halpern et al. (2004), van Benthem and Pacuit (2006), van Benthem et al. (2009)), the protocol is only implicitly represented, for example, with statements of the form "$F\varphi$" meaning that "φ is true at some moment in the future (after the agents perform actions consistent with the protocol)."

In this paper, we develop a logical framework where protocol(s) are "first-class citizens" (cf. van Benthem 2001a). This provides a local perspective where simple protocols can be combined to construct more complex ones. Thus, we drop the assumption that there is a single, fixed protocol and consider situations where the protocol is created "as needed." A number of authors have suggested different variations of *propositional dynamic logic* (PDL) to reason about protocols, or *strategies*, from this local, "constructive" point of view (for example, see Fagin et al. 1995, van Benthem 2001a, van Benthem and Pacuit 2006, van Eijck et al. 2009, Wang 2010). The idea is that PDL-action expressions explicitly describe different protocols. Under this interpretation, the PDL formula $[\pi]\varphi$ has the interpretation "φ is guaranteed to be true by following the protocol π."

[3]Cf. van der Meyden (1996), where the models are generated by unfolding some multi-agent finite state machine.

Here, "following the protocol π" means that agent(s) makes choices so that the resulting sequence of events matches π.

We start with a single agent who, in each possible state, can choose from a finite set of actions (the actions she "can" perform in the sense of points 1 and 2 above). Each action corresponds to a (possibly nondeterministic) transition from the current state to a new state, and there may be different actions available at different states. In other words, we assume that the agent is in a *labeled transition system*, which we call an **arena**. The arena describes the actions that are available at each state and the possible consequences of each action. The following is an example of an arena:



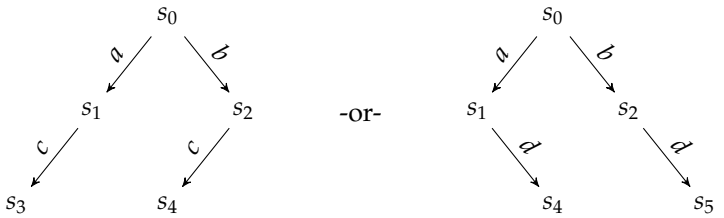
A **protocol** is a tree with labels from the (finite) set of possible actions. We are interested in what properties the agent(s) can *guarantee* will be true by *adopting a given protocol*. The idea is that adopting a protocol at a state restricts the paths that the agent will follow from that state. In general, adopting a protocol does not commit the agent(s) to a single course of action, but, rather, focuses the agent's(s') attention on the "relevant" decision problems. Thus, "adopting a protocol" simply amounts to "committing to a *plan*," something that is crucial for an autonomous (rational) agent. In his influential book, Michael Bratman (1987) argues, *inter alia*, that

> plans help make deliberation tractable for limited beings like us. They provide a clear, concrete purpose for deliberation, rather than merely a general injunction to do the best. They narrow the scope of the deliberation to a limited set of options. And they help answer a question that tends to remain unasked within traditional decision theory, namely; where do decision problems come from?
>
> (Bratman 1987, pg. 33)

One contribution of our paper is to explore the conditions under which agent(s) can engage in such (future-directed) planning (cf. Cohen and Levesque 1990,

Meyer et al. 1999). We focus on *structural properties* of the interactive situation (i.e., what the agents *can* do) and what the agents "know" about the decision problems they face. We leave for future work how to incorporate the agents' *motivating attitudes* (e.g., desires, goals, wishes) into our logical analysis. Thus, we focus on when the agent(s) *can* (implicitly or explicitly) agree to adopt a protocol, or commit to a plan, instead of why the agent(s) would *want* to agree to a protocol, or plan.

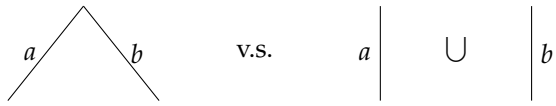
Our first observation is that it is important to interpret the PDL actions expressions over *finite trees* rather than *paths*. In other words, our basic actions expressions denote finite trees instead of the usual one-step actions (cf. Ramanujam and Simon 2009). For example, suppose that the agent is in state s_0 in the above arena and consider the protocol "either choose c or choose d." This protocol gives only partial information about what actions to follow at a given state (e.g., the protocol does not offer any advice about what to do at s_0). This protocol can be described by the PDL expression $(a \cup b); c \cup (a \cup b); d$. Note that every path in the above arena is consistent with this protocol, so we can say that this protocol is *enabled* at s_0. However, as Johan van Benthem (2010) points out, this way of thinking about the protocol misses a crucial point: The agent must commit to do either c or d *independent* of which action is chosen at state s_0. In other words, by committing to this protocol (at s_0), the agent must choose between the following two restrictions on future choices:



This distinction is not important if we are interested in only the states that can result by following this protocol—in this case, $\{s_3, s_4\} \cup \{s_4, s_5\}$. However, it is becomes important when constructing complex plans from simpler ones using the regular operations of PDL (union $\cup$, concatenation ; and Kleene star *) or if an agent conditions on the plans of another agent (or her future self).

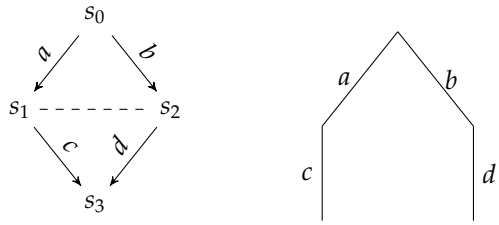
An interesting feature of allowing branching in atomic programs is that we can represent a choice between a and b in two different ways. The picture on the left denotes the atomic tree consisting of two branches, one labeled with a and the other with b. The picture on the right is a complex program built using

the union operator from two atomic trees, each containing only one branch.



These two programs have very different interpretations corresponding to different ways of understanding what it means for an agent to commit the plan: do a or do b. On the first interpretation, the agent commits to choosing between actions a or b *when the time comes* (possibly ignoring the other options that may be available to the agent at that moment). On the second interpretation, the agent must choose between two future courses of actions: doing a or doing b. The point is that a and b each may lead to a different set of states.

Our main contribution in this paper is to analyze different ways in which a protocol "can" be adopted (by either a single agent or a group of agents) taking each agent's *point of view* into account. Since we assume that actions may be nondeterministic, there may be many ways in which a protocol can be "realized" at a position in an arena. This creates uncertainty for the agent since, in general, she may not know which state results from a particular action. However, there may be other sources of imperfect information. For example, the agent may have only limited memory or observational power, or the agent may be uncertain about the exact "starting position" or initial state of the situation. Thus, at certain positions in the arena, for whatever reason, it may appear to the agent that she is in a different position or set of positions. For example, consider the following situation where the agent cannot distinguish between nodes s_1 and s_2 and the protocol pictured to the right (do a followed by c or do b followed by d):



This protocol is clearly enabled in the situation without the uncertainty relation between s_1 and s_2. However, in the above situation at s_0, the agent cannot agree

to "*knowingly*" follow the protocol since she is uncertain about the actions that are available at states[4] s_1 and s_2.

2 Our framework

We assume the reader is familiar with standard definitions of trees and arenas (i.e., labeled transition systems or Kripke models). A **protocol** is a finite labelled tree. Let Σ be a finite set whose elements are called **actions**. A Σ-labelled (finite) tree T is a tuple $(S, \{\Rightarrow_a\}_{a\in\Sigma}, s_0)$ where S is a (finite) set of nodes, $s_0 \in S$ is the root and for each $a \in \Sigma$, $\Rightarrow_a \subseteq S \times S$ is the edge relation satisfying the usual properties. For a node $s \in S$, let $\mathcal{A}(s) = \{a \in \Sigma \mid \exists s' \in S \text{ where } s \Rightarrow_a s'\}$ denote the set of *actions available at* s.

We formally model an interactive (or decision-theoretic) situation in a standard way as a labelled transition system which we call **arenas**: Let W be a nonempty finite set, whose elements are called **positions** or **states**, and Σ a finite set of basic actions. An **arena** is a structure $\mathcal{G} = (W, \{\Rightarrow_a\}_{a\in\Sigma})$ where for each $a \in \Sigma$, $\Rightarrow_a \subseteq W \times W$. Following standard notation, we write $w \Rightarrow a\, v$ if $(w, v) \in \Rightarrow a$. The above notation for available actions and paths are readily applied to finite arenas.

A protocol or plan *restricts* the available choices for the agent(s). Intuitively, if an agent agrees to follow a finite protocol, then she agrees to restrict her choices to all and only those actions compatible with the protocol. Of course, not all protocols *can* be followed in any situation. This leads us to the key notion of a protocol being **enabled** at a state u in an arena. If there is no uncertainty in the arena, then the formal definition of a protocol being enabled is completely straightforward: a protocol T is enabled at u in $\mathcal{G}$ if T can be embedded in the unwinding of u.

Notation. For a node $s \in S$, let $\mathcal{A}(s) = \{a \in \Sigma \mid \exists s' \in S \text{ where } s \Rightarrow_a s'\}$ denote the set of *actions available at* s. A node s is called a *leaf node* if $\mathcal{A}(s) = \emptyset$, and the set of all leaf nodes in the tree is denoted by *frontier*(T). For a set X and a finite sequence $\rho = x_1 x_2 \dots x_m \in X^*$, let $last(\rho) = x_m$ denote the last element in this sequence and $first(\rho) = x_1$ the first element. We extend this notion to a set $Y \subseteq X^*$ as $last(Y) = \{x \mid \exists \rho \in Y \text{ with } last(\rho) = x\}$. The following definition is standard: A **path** in the tree $T = (S, \{\Rightarrow_a\}_{a\in\Sigma}, s_0)$ is an alternating sequence of nodes and actions $\rho = s_0 a_0 s_1 a_1 \cdots a_{k-1} s_k$ satisfying the following condition: for

[4] Alternatively, we can say that the agent forgets at state s_1 (and s_2) the choice that was made at state s_0.

all $j : 0 \leqslant j < k$, we have $s_j \Rightarrow_{a_j} s_{j+1}$. The **length** of a path ρ, denoted $len(\rho)$, is the number of actions appearing in ρ. A path ρ is **maximal** in T if $first(\rho) = s_0$ and $\mathcal{A}(last(\rho)) = \emptyset$. Let $Paths(T)$ denote the set of all maximal paths in T. For $\rho = s_0 a_0 s_1 a_1 \dots s_k$, let $head(\rho) = a_0$ and $tail(\rho) = s_1 a_1 \dots s_k$.

In some cases, it is convenient to define a path as a sequence of states (or actions). For example, we say a sequence of *states* $\sigma = s_0 s_1 \cdots s_k$ is a **path of states** if there are actions $a_0, \dots, a_{k-1}$ such that $s_0 a_0 s_1 a_1 \cdots a_{k-1} s_k$ is a path (define a **path of actions** similarly). We can use these definitions to define the **height** of a finite tree T (the length of the longest path): $height(T) = \max\{len(\rho) \mid \rho \in Paths(T)\}$. Note that the above labeled trees may be *nondeterministic* since two edges from the same node can have the same label (i.e., there may be distinct nodes s, s' and s'' such that $s \Rightarrow_a s'$ and $s \Rightarrow_a s''$). However, if the tree is intended to represent a *protocol* or *plan* that an agent has committed to follow, then it is natural to restrict attention to *deterministic* trees:

Finally, let $\mathcal{G}$ be an arena. The unwinding at state u in $\mathcal{G}$ is a tree which is denoted by T_u. If T is a basic protocol, then we denote that T is **enabled at** u by $enabled(t, u)$. The formal details of these notions are standard, so we do not include them here (see the full paper (Pacuit and Simon 2010) for details).

Intuitively, if a protocol T is enabled at a state u in an arena $\mathcal{G}$, then it is (physically, objectively) *possible* for the agent to *agree* to follow T. Of course, this does not necessarily mean that the agent *knows* (or *believes*) she can follow T, *wants* to follow T or it is in the agent's interests for follow T. Our main goal in this paper is to explore a different sense in which a protocol is "possible" taking into account the agent's *point-of-view*. Our first task is to extend the definition of an arena with an explicit representation of the agent's "point-of-view" at each position in the arena. As is standard in the epistemic logic literature, we use a relation on the set of states in an arena to represent this uncertainty of the agent about her position in the arena.

Definition 2.1 (Arena with Imperfect Information)**.** An **arena with imperfect information** is a structure $\mathcal{G}^I = (W, \{\Rightarrow a\}_{a \in \Sigma}, \rightsquigarrow)$ where $(W, \{\Rightarrow a\}_{a \in \Sigma})$ is a finite arena and $\rightsquigarrow \subseteq W \times W$.

For each position u in an arena, let $\mathcal{I}(u) = \{w \mid u \rightsquigarrow w\}$ be the agent's "*point-of-view*". The above models do not impose any structural properties on the action and $\rightsquigarrow$ relations. However, a number of properties discussed in the literature are natural in many situations. Suppose that the agent is in position w but "thinks" she is in position v (i.e., $w \rightsquigarrow v$), and consider an action $a \in \mathcal{A}(w) \cap \mathcal{A}(v)$. In this case, the agent is aware that she can do a and will not

fail. Furthermore, unless there is a "miracle," doing action a should not remove the agent's "uncertainty" (e.g., the $\rightsquigarrow$ relation). Formally,

- **No Miracles**: For all $a \in \Sigma$ and all $w, v, w', v' \in W$, if $w \rightsquigarrow v$, $w \Rightarrow_a w'$, and $v \Rightarrow_a v'$, then $w' \rightsquigarrow v'$.

Imposing no miracles means that the basic actions are assumed to be "uninformative". No miracles covers the situation when $a \in \mathcal{A}(w) \cap \mathcal{A}(v)$ (recall that $\mathcal{A}(w)$ is the set of actions available at w). The remaining interesting situations are when an action a is available only in one of the states. First, if $a \in \mathcal{A}(w)$, but $a \notin \mathcal{A}(v)$, then the agent does not realize that a is actually available. Second, if $a \in \mathcal{A}(v)$, but $a \notin \mathcal{A}(w)$, then the agent believes that she can do a, but will fail[5] if she attempts to execute this action. Each of these situations is covered by the following two properties:

- **Success**: If $w \rightsquigarrow v$, then $\mathcal{A}(v) \subseteq \mathcal{A}(w)$.
- **Awareness**: If $w \rightsquigarrow v$, then $\mathcal{A}(w) \subseteq \mathcal{A}(v)$.

Of course, if $\rightsquigarrow$ is symmetric, then these properties are equivalent and we have $\mathcal{A}(w) = \mathcal{A}(v)$ provided $w \rightsquigarrow v$. These properties address the relationship between the actions available at the current state (which the agent may not have access to) and the actions available at states the agent considers "possible" (via $\rightsquigarrow$). The next property focuses on the relationship between the actions available at the set of states the agent considers "possible." If $w \rightsquigarrow v$ and $w \rightsquigarrow v'$, then the agent may find herself in either v or v' and so should face the same decision problem:

- **Certainty of available actions**: If $w \rightsquigarrow v$ and $w \rightsquigarrow v'$, then $\mathcal{A}(v) = \mathcal{A}(v')$.

Of course, these properties are all equivalent in the important special case when the agent's information relation ($\rightsquigarrow$) is an equivalence relation (a common assumption in the epistemic logic and game theory[6] literature). This special case is particularly interesting since it helps position our work within the broad literature using various combinations of modal logics to reason about game/decision-theoretic situations (cf. Lorini et al. 2009, van Benthem 2001).

[5]Note that we do not address in this paper what happens (from the agent's point of view) if she tries to do an action a that is not actually available (i.e., the agent *attempts* action a). This interesting situation will be addressed in future work. See Lorini and Herzig (2008) for a very interesting discussion relevant to this situation.

[6]Of course, game theorists tend to focus on arenas that are themselves *trees*—i.e., extensive games with imperfect information.

However, as discussed above, this is an *objective* notion from the modeler's point of view that does not take into account that the agents may be imperfectly informed about their "location" in the arena. What we need is a *subjective* version of being enabled. The idea to ensure *at each step* that we take into account all and only the positions that the agent has access to via the $\rightsquigarrow$ relation. Intuitively, a protocol T is *subjectively enabled* at a position u in an arena with imperfect information if:

(1) the agent is *certain that* T is enabled (for all $v \in \mathcal{I}(u)$, T is enabled at v); and

(2) the agent will be certain that she is, in fact, following the protocol at *every stage* of the protocol.

This second point is important as there is a difference between "knowing that a protocol is enabled" and "being able to *knowingly follow* a protocol."[7] This difference is crucial for an agent contemplating committing to a long-term plan.[8] Thus, our definition must take into account the forest $\{T_v \mid v \in \mathcal{I}(u)\}$ *for every position u not ruled out by the protocol*.

Recall that *enabled*(T, u) is true if there is an embedding of T into T_u. We have to complicate this simple picture in the presence of imperfect information. We start by stating the most general definition and then show how to simplify it in the presence of the structural assumptions discussed above (e.g., assuming $\rightsquigarrow$ is an equivalence relation). First of all, note that in arenas with imperfect information, the restriction of a protocol T is not a tree, but, rather, a *forest* (possibly containing trees of different heights). Thus, we need to introduce notation for forests in an arena. Let $\mathcal{G}$ be an arena (with imperfect information). First, recall that the notion of a path applies to arenas and, by assumption, the last element of a path is always a state. We say that a path ρ is an *initial segment* of ρ' if ρ' is ρ followed by a possibly empty path. Formally, $\rho = w_0 a_0 \cdots a_{k-1} w_k$ is an initial segment of ρ' if there is an $i \geq 0$ such that $\rho' = w_0 a_0 \cdots a_{k-1} w_k a_{k+1} \cdots a_{k+i-1} w_{k+i}$. Given a set of paths X that is closed under initial segment, we define an edge relation in the obvious way: $\rho \Rightarrow_a^X \rho'$ iff $\rho = w_0 a_0 \cdots a_{k-1} w_k$ and $\rho' = w_0 a_0 \cdots a_{k-1} w_k a w$. A set of paths X from an arena

[7] See Broersen (2008) for a discussion related to this point.

[8] After all, an agent cannot commit to a temporally-extended plan if she is certain now that she will not be able to choose in a way that is consistent with that plan. Of course, this does not preclude the possibility that the agent may need to revise or drop her plan *even after committing to it* (perhaps because she learned that the plan is no longer feasible). See Icard et al. (2010) for a complete discussion.

$\mathcal{G}$ that is closed under initial segment is called a **forest in** $\mathcal{G}$ if $\{\Rightarrow_a^X\}_{a\in\Sigma}$ satisfies the properties 1, 2 and 3 in the definition of a tree given above.

It is not hard to see that if a protocol T is enabled at u, then the restriction of T at u gives us a forest X with each path in X is associated with a node in T. Generalizing to situations with imperfect information, we may need to associate more than one path with a node in T. Thus, we define the restriction of T in an arena with imperfect information to be a forest X and function mapping paths in X onto nodes in T:

Definition 2.2 (Subjective Restriction)**.** Let $\mathcal{G}^I = (W, \{\Rightarrow a\}_{a\in\Sigma}, \rightsquigarrow)$ be an arena with imperfect information, $u \in W$ and $T = (S, \{\Rightarrow_a\}_{a\in\Sigma}, s_0)$ a protocol. The **subjective restriction** of T in $(\mathcal{G}^I, u)$, denoted $(\mathcal{G}^I, u)\restriction_s T$, is a pair (X, f) where X is a forest in $\mathcal{G}^I$ and f is a function from X onto S. Both X and f are defined inductively as follows:

0. $X_0 = \mathcal{I}(u)$ ($v \in X_0$ is understood as a one-element sequence) and for all $v \in X_0$, set $f_0(v) = s_0$

n. Suppose X_n and f_n have been constructed, for each $\rho \in S_n$, for all $a \in \mathcal{A}(f_n(\rho))$, let $Y_a = \{\rho aw \mid last(\rho) \Rightarrow a\, w \text{ in } \mathcal{G}^I\} \cup \bigcup\{\mathcal{I}(w) \mid last(\rho) \Rightarrow a\, w \text{ in } \mathcal{G}^I\}$. Define
$$X_{n+1} = X_n \cup \bigcup_{a\in\mathcal{A}(f_n(\rho)),\rho\in X_n} Y_a$$
Let f_{n+1} extend f_n such that for each new node $\rho aw \in Y_a$, set $f_{n+1}(\rho aw) = s'$ where $f_n(\rho) \Rightarrow_a s'$ in[9] T.

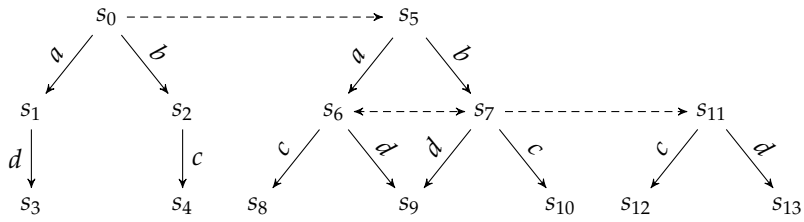
Let $X = X_{height(T)}$ and $f = f_{height(T)}$. Finally, define the **frontier of** $(\mathcal{G}^I, u)\restriction_s T$ as follows: $frontier((\mathcal{G}^I, u)\restriction_s T) = \{last(\rho) \in W \mid \mathcal{A}(f(\rho)) = \emptyset\}$.

Define the actions available at a path in a forest as follows: suppose that X is a forest and $\rho \in X$ and define $\mathcal{A}(\rho) = \{a \in \Sigma \mid \text{ there is a } \rho' \in X \text{ such that } \rho \Rightarrow_a^X \rho'\}$.

Definition 2.3 (Subjectively Enabled)**.** A protocol T is **subjectively enabled** at u in $\mathcal{G}^I = (W, \Rightarrow, \rightsquigarrow)$, denoted $s\text{-}enabled(T, (\mathcal{G}^I, u))$, if the structure $(\mathcal{G}^I, u)\restriction_s T = (X, f)$ satisfies the condition $\forall \rho \in X, \mathcal{A}(\rho) = \mathcal{A}(f(\rho))$.

Notice that without additional structural assumptions on $\rightsquigarrow$, a protocol being subjectively enabled does *not* imply that the protocol is enabled. For example, consider the arena below and the protocol discussed in the introduction: "either do a followed by c or do b followed by d." This protocol is subjectively enabled but not enabled at state s_0.

[9] Since T is deterministic, f_{n+1} is well defined.



(*Note that the protocol is still subjectively enabled if we impose the no miracle property, which would add a number $\rightsquigarrow$ edges.*)

We conclude this section with two observations. The first is that in situations of *perfect information*, subjectively enabled is equivalent to enabled:

Proposition 1. *Suppose that $\mathcal{G}^I = (W, \{\Rightarrow a\}_{a\in\Sigma}, \rightsquigarrow)$ satisfies the property that for all $w \in W$, $\mathcal{I}(w) = \{w\}$. Then, for any protocol T and state $w \in W$, T is enabled at w in $(W, \{\Rightarrow a\}_{a\in\Sigma})$ iff T is subjectively enabled at w in $\mathcal{G}^I$.*

The proof follows by unpacking the definitions and is left to the reader. Additional structural properties can further simplify the definition of subjectively enabled. We have already remarked that a protocol being "subjectively enabled" at a state w is, in general, not equivalent to the agent knowing that the protocol is enabled at w (i.e., the protocol is objectively enabled at every state in $\mathcal{I}(w)$) A simple argument shows that these notions coincide when the agent is certain of her available actions and the actions are not informative:

Lemma 1. *Suppose $\mathcal{G}^I = (W, \{\Rightarrow a\}_{a\in\Sigma}, \rightsquigarrow)$ satisfies certainty of actions and no miracles. Then, the agent* knows that T is enabled *at u iff T is subjectively enabled at u (i.e.,* s-enabled$(T, (\mathcal{G}^I, u))$ *is true).*

2.1 What can be achieved with protocols?

An arena with imperfect information describes what *can happen* in an interactive situation both objectively (from the modeler's point-of-view) and subjectively (from the agent's point-of-view via the $\rightsquigarrow$ relations). That is, they describe both what is physically possible for the agent to do and what she thinks she can do in an interactive situation.

Committing to a basic protocol T *restricts* the choices available to the agent, but there is a trade-off: it also *increases* the ability of the agent to *guarantee* that certain propositions are true. Formally, each basic protocol (which is a finite

tree) is associated with a set of states X (the *frontier* of T in an arena). These are the states that the agent can "force" the situation to end up in by making choices consistent with the protocol. There are a number of ways to make precise what it means for an agent to "guarantee" that some proposition is true because she adopts the protocols T. One options is to see what is true no matter what the agent does as long as it is consistent T. A second option recognizes that T still represents choices for the agent which will be settled in the course of the interaction. In this case, we are interested in what the agent can force by doing something consistent with T. The situation is even more interesting when the agent commits to a complex protocol. If the protocol involves the operators $\cup$ or Kleene star then the agent first must choose which set of states she wants to have the ability to force. For example, consider the protocol $T_1 \cup T_2$, in order to commit to this protocol the agent must choose which of the two basic protocols to follow. More generally, given a complex protocol π, the agent must first decide both *how* to go about adopting π then make her choices "in the moment" consistent with this plan.

This discussion suggests that our basic modality will be interpreted as a sequence of *two* quantifiers (each corresponding to the different "types" of decisions the agent makes when committing to a protocol). This is familiar from other modal logics of ability (eg., STIT) and game logics. Of the four possible combinations of quantifiers, we take the following two as primitive (corresponding to $\exists\forall$ and $\exists\exists$ respectively):

- $\langle\pi\rangle^{\forall}\alpha$: By adopting the protocol π, α is guaranteed to be true.
- $\langle\pi\rangle^{\exists}\alpha$: By adopting the protocol π, the agent can do something consistent with the protocol that will make α true.

As usual, the remaining two possible combinations of quantifiers are dual to these. We take "adopting a protocol" to mean that the agent decides how to follow the protocol (so an existential quantifier over the different sets of states the agent can force). The second quantifier is over the different ways that the agent actually implements the protocol. These notions are objective since they do not take into account the fact that the agent may be imperfectly informed about her current position in the arena. This suggests the following "epistemized" versions of the above operators:

- $\langle\pi\rangle^{\Box}\alpha$: By *agreeing* to adopt the protocol π, the agent is certain that α is guaranteed to be true.

- $\langle\pi\rangle^{\diamond}\alpha$: By *agreeing* to adopt the protocol π, the agent can "knowingly" do something consistent with the protocol that will make α true.

2.2 Epistemic protocol logic

Thus far, we have focused only on *basic protocols*. It is convenient to give an explicit syntax for describing basic protocols.

Definition 2.4 (Syntax for Protocols)**.** Let $\mathcal{V}$ be a countable set of node variables. A **protocol expression** is inductively defined as follows:

- For each $x \in \mathcal{V}$, (x) is a protocol expression.
- Suppose that $J = \{a_1, \ldots, a_m\}$ is a set of (distinct) actions and for each a_i we have a (unique) protocol expression t_{a_i}. Then,

$$(x, a_1, t_{a_1}) + \cdots + (x, a_m, t_{a_m})$$

  is a protocol expression where x is a new variable not appearing in t_{a_i}.

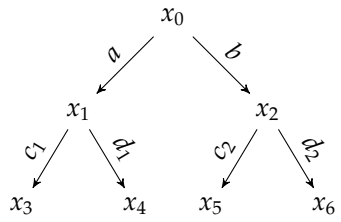
Let $\mathcal{P}(\mathcal{V})$ denote the set of protocol expressions.

The idea is that the expression (x, a, t_a) denotes the subtree where x is the root and there is an a-edge from x to the subtree described by t_a. Note that this syntax generates only deterministic trees (i.e., basic protocols) since each action a in a protocol expression is associated with only one subtree. Of course, there are other ways to syntactically describe finite trees, but the particular choice of syntax is not crucial for our analysis. The important point is that each syntactic expression $t \in \mathcal{P}(\mathcal{V})$ corresponds to a finite tree T_t:

Definition 2.5 (Interpretation of Protocol Expressions)**.** Given $t \in \mathcal{P}(\mathcal{V})$, we can inductively define the **basic protocol** T_t **generated by** t as follows:

- if $t = (x)$, then let $T_t = (S_t, \Rightarrow_t, s_x)$ where $S_t = \{s_x\}$ and $\Rightarrow_t = \emptyset$.
- if $t = (x, a_1, t_{a_1}) + \cdots + (x, a_k, t_{a_k})$, then inductively we have trees $T_1, \ldots T_k$ where for $j : 1 \leqslant j \leqslant k$, $T_j = (S_j, \Rightarrow_j, s_j)$. Define $T_t = (S_t, \Rightarrow_t, s_x)$ where s_x is a new state and
  - $S_t = \{s_x\} \cup S_{T_1} \cup \ldots \cup S_{T_k}$.
  - $\Rightarrow_t = (\bigcup_{j=1,\ldots,k} \Rightarrow_j) \cup \{s_x \Rightarrow_{a_j j} s_j \mid 1 \leqslant j \leqslant k\}$. ◂

For $t \in \mathcal{P}(\mathcal{V})$, we often abuse notation and identify t with T_t. The following example illustrates the above construction:



The syntactic representation of this tree using Definition 2.4 is:

- $t = (x_0, a, t_1) + (x_0, b, t_2)$ where
  - $t_1 = (x_1, c_1, (x_3)) + (x_1, d_1, (x_4))$ and
  - $t_2 = (x_2, c_2, (x_5)) + (x_2, d_2, (x_6))$.

The next step is a syntax for describing *complex protocols*. To keep things simple, we focus on the regular operations familiar from action logics such as PDL: Let Σ be a finite set of basic actions, and define Γ to be the smallest set of expressions generated by the following grammar:

$$t \mid \pi_1; \pi_2 \mid \pi_1 \cup \pi_2 \mid \pi^*$$

where $t \in \mathcal{P}(\mathcal{V})$ is a basic protocol (using actions from Σ). Note that we do not include tests in our language. Adding tests raises a number of interesting issues (many have been extensively discussed in the literature on knowledge programs Fagin et al. 1995; 1997); however, we leave this extension for future work.[10] We can easily adopt the standard interpretation of these operations to our setting:

(1) $\pi_1; \pi_2$ is the protocol where the agent first adopts the protocol π_1 and then (no matter what happens) adopts the protocol π_2;

(2) $\pi_1 \cup \pi_2$ is the protocol where the agent must first choose which of the two protocols to adopt; and

(3) π^* is the protocol where the agent continues with protocol π any finite number of times (including zero).

Of course, there may be other natural operations in this context, such as "merging"[11] or "revising" (cf. Icard et al. 2010).

[10] Note that there is nothing inherently difficult about adding tests to our language; and, indeed, the results in this paper can be adapted to this situation. We do not include them here to simplify the setting and focus on issues that are orthogonal to issues that are relevant when tests are in the language.

[11] Yanjing Wang (2010) has an extensive discussion in his dissertation (using PDL).

Let At be a countable set of atomic propositions and Γ a set of protocol expressions as defined in Definition 2.5 (based on basic actions Σ). The **epistemic protocol language** is the smallest set $\mathcal{L}_{EPL}$ of formulas generated by:

$$p \in \mathsf{At} \mid \neg\alpha \mid \alpha_1 \vee \alpha_2 \mid \Box\alpha \mid \langle\pi\rangle^{\exists}\alpha \mid \langle\pi\rangle^{\forall}\alpha \mid \langle\pi\rangle^{\Box}\alpha \mid \langle\pi\rangle^{\Diamond}\alpha$$

where $\pi \in \Gamma$. As a convention we use $\top = p \vee \neg p$. We also define $\Diamond\alpha = \neg\Box\neg\alpha$, $[\pi]^{\exists}\alpha = \neg\langle\pi\rangle^{\forall}\neg\alpha$, $[\pi]^{\forall}\alpha = \neg\langle\pi\rangle^{\exists}\neg\alpha$, $[\pi]^{\Diamond}\alpha = \neg\langle\pi\rangle^{\Box}\neg\alpha$ and $[\pi]^{\Box}\alpha = \neg\langle\pi\rangle^{\Diamond}\neg\alpha$. We discussed the four protocol modalities above. The remaining modality $\Box$ quantifies over states accessible (in one step) via the $\rightsquigarrow$ relation. Thus, it describes what is true from the agent's point of view. As usual, models are arenas with valuation functions:

Definition 2.6 (Model)**.** Let $\mathcal{G}^I = (W, \{\Rightarrow a\}_{a\in\Sigma}, \rightsquigarrow)$ be an arena with imperfect information. A **model** based on $\mathcal{G}^I$ is a structure $(W, \{\Rightarrow a\}_{a\in\Sigma}, \rightsquigarrow, V)$ where and $V : \mathsf{At} \to 2^W$ a valuation function.

Before defining truth in a model, we must "interpret" complex protocols. The idea is to associate with each protocol π the collection of states that the agent can force by following π. Formally, we define sets $R^Q_\pi \subseteq W \times 2^W$ for $Q \in \{\exists, \forall, \Box, \Diamond\}$ by induction on the structure of π. We start with the atomic protocols.

Atomic Protocols. For an atomic protocol expressions t, and $Q \in \{\exists, \forall, \Box, \Diamond\}$, we define the relation $R^Q_t \subseteq W \times 2^W$ as follows:

- $R^{\mathfrak{F}}_t = \{(u, X) \mid \mathit{enabled}(T_t, u)$ and $\mathit{last}(\mathit{frontier}(T_u\lceil T_t)) = X\}$ (for $\mathfrak{F} \in \{\exists, \forall\}$).
- $R^{\Box}_t = \{(u, X) \mid \mathit{s\text{-}enabled}(T_t, u)$ and $\mathit{last}(\mathit{frontier}((\mathcal{G}, u)\restriction_s T_t)) = X\}$.

The definition of $R^{\Diamond}_t$ is more complicated. The issue is that, in this case, the way the agent implements the protocol must take into account the agent's imperfect information. This suggests the following notion: given a path $\rho = s^0_t a_0 s^1_t \ldots s^k_t \in \mathit{Paths}(t)$, the **subjective path** defined by ρ on the structure $(\mathcal{G}, u)\restriction t = (S, \Rightarrow, f)$ is the sequence $\mathfrak{S}(\rho, u) = Z_0 Z_1 \ldots Z_k$ where for all $j : 0 \leqslant j \leqslant k$, $Z_j = \{s \in S \mid f(s) = s^j_t\}$. We now have

- $R^{\Diamond}_t = \{(u, X) \mid \mathit{s\text{-}enabled}(T_t, u)$ and $\exists\rho \in \mathit{Paths}(T_t)$ with $\mathfrak{S}(\rho, u) = Z_0 Z_1 \ldots Z_k$ and $X = Z_k\}$.

Composition.

- $R^{\exists}_{\pi_1;\pi_2} = \{(u, X) \mid \exists Y \subseteq W$ such that $(u, Y) \in R^{\exists}_{\pi_1}$ and $\exists v_j \in Y$ such that $(v_j, X) \in R^{\exists}_{\pi_2}\}$.
- for $\mathfrak{F} \in \{\forall, \Box, \Diamond\}$,
  - $R^{\mathfrak{F}}_{\pi_1;\pi_2} = \{(u, X) \mid \exists Y = \{v_1, \ldots, v_k\}$ such that $(u, Y) \in R^{\mathfrak{F}}_{\pi_1}$ and $\forall v_j \in Y$, there exists $X_j \subseteq X$ such that $(v_j, X_j) \in R^{\mathfrak{F}}_{\pi_2}$ and $\bigcup_{j=1,\ldots,k} X_j = X\}$.

Note that in the definition above, we can assume the set Y is finite since our models are finitely branching. The definition of union and Kleene star is standard (though some care must be taken in the latter case to use a fixed-point definition):

Union. For $Q \in \{\exists, \forall, \Box, \Diamond\}$, $R^{Q}_{\pi_1 \cup \pi_2} = R^{Q}_{\pi_1} \cup R^{Q}_{\pi_2}$.

Iteration.

- $R^{\exists}_{\pi^*} = \bigcup_{n \geqslant 0} (R^{\exists}_{\pi})^n$.

For $Q \in \{\forall, \Box, \Diamond\}$, it is tempting to define iteration as $R^{Q}_{\pi^*} = \bigcup_{n \geqslant 0} (R^{Q}_{\pi})^n$. However, this definition does not give the intended interpretation of the Kleene star operator. To see this, consider the simple tree t consisting of a root and two outgoing edges a and b. Intuitively, the above definition would force all the branches of t^* to be of the same depth. This also illustrates the underlying difference between our approach and that of standard dynamic logic: Sequential composition in our setting is defined over trees rather than over paths. The semantics of Kleene star, thus, needs to be defined with respect to a least fixed-point operator. We formalize this as follows: Let $\cdot$ be a binary operator over $W \times 2^W$, which is defined as:

- $R_1 \cdot R_2 = \{(u, X) \mid \exists w_1, Y_1, \ldots, w_k, Y_k$ with $(u, \{w_1, \ldots, w_k\}) \in R_1$, $\forall j, (w_j, Y_j) \in R_2$ and $X = \bigcup_j Y_j\}$.

for all $R_1, R_2 \subseteq W \times 2^W$.

Given a $Z \subseteq W \times 2^W$, let F_Z be the operator over the domain $W \times 2^W$ defined as $F_Z(R) = R_\top \cup Z \cdot R$ where $R_\top = \{(u, \{u\}) \mid u \in W\}$. Observe that the operator $\cdot$ is monotonic in the following sense: If $R_1 \subseteq R_2$, then $R_0 \cdot R_1 \subseteq R_0 \cdot R_2$. This also implies that F_Z is monotonic for every $Z \subseteq W \times 2^W$. Thus, by the Knaster-Tarski theorem we have that for every Z, the least fixed-point (*LFP*) of F_Z exists. $LFP(F_Z)$ can be computed as the limit of the following sequence of

partial solutions: $R_0 = R_\top, R_{j+1} = F_Z(R_j)(= R_\top \cup Z \cdot R_i)$ and $R_\lambda = \cup_{\nu<\lambda} R_\lambda$ for a limit ordinal λ. For $Q \in \{\forall, \Box, \Diamond\}$, we define:

- $R^Q_{\pi^*} = LFP(F_{R^Q_\pi})$.

We are now in a position to formally define truth in a model:

Definition 2.7 (Truth)**.** Let $M = (W, \Rightarrow, \rightsquigarrow, V)$ be a model with position u. A formula $\alpha \in \mathcal{L}_{EPL}$ is **true** at state u in M (denoted $M, u \models \alpha$) is defined as follows:

- $M, u \models p$ iff $p \in V(u)$
- $M, u \models \neg\alpha$ iff $M, u \not\models \alpha$
- $M, u \models \alpha_1 \vee \alpha_2$ iff $M, u \models \alpha_1$ or $M, u \models \alpha_2$
- $M, u \models \Box\alpha$ iff for all w such that $u \rightsquigarrow w$ we have $M, w \models \alpha$
- $M, u \models \langle\pi\rangle^\exists\alpha$ iff $\exists(u, X) \in R^\exists_\pi$, $\exists w \in X$ such that $M, w \models \alpha$
- $M, u \models \langle\pi\rangle^\forall\alpha$ iff $\exists(u, X) \in R^\forall_\pi$ such that $\forall w \in X$ we have $M, w \models \alpha$
- $M, u \models \langle\pi\rangle^\Box\alpha$ iff $\exists(u, X) \in R^\Box_\pi$ such that $\forall w \in X$ we have $M, w \models \alpha$
- $M, u \models \langle\pi\rangle^\Diamond\alpha$ iff $\exists(u, X) \in R^\Diamond_\pi$ such that $\forall w \in X$ we have $M, w \models \alpha$

where for $Q \in \{\exists, \forall, \Box, \Diamond\}$, $R^Q_\pi \subseteq W \times 2^W$ is defined above. The logical notions *satisfiability* and *validity* are defined as usual.

The first technical contribution of this paper is a sound and (weakly) complete axiom system (in the language $\mathcal{L}_{EPL}$) for the class of all arenas with imperfect information. A straightforward consequence of this completeness proof is decidability of the satisfiability problem, which we discuss below.

The axiomatization and completeness proof extends the one found in Ramanujam and Simon (2009) to situations with imperfect information. In this section, we present this axiom system and discuss the proof (details can be found in the full paper). First of all, note that the language $\mathcal{L}_{EPL}$ extends the standard PDL language: Let e_a denote the tree $e_a = (x, a, y)$ with a single a-edge, and define for each $a \in \Sigma$, $\langle a\rangle\alpha = \langle e_a\rangle^\exists\alpha$. Given the semantics defined above (Definitions 2.6 and 2.7), we have the standard interpretation for $\langle a\rangle\alpha$: $\langle a\rangle\alpha$ holds at a state u iff there is a state w such that $u \xrightarrow{a} w$ and α holds at w.

A key observation is that whether a protocol t is (subjectively) enabled can be described by a standard PDL formula. Formally, for each protocol T, let $t^{\surd}$ be a formula that is intended to denote that the tree structure t is enabled. This is defined inductively on the structure of t as:

- if $t = (x)$, then $t^{\surd} = \top$.
- if $t = (x, a_1, t_{a_1}) + \ldots + (x, a_k, t_{a_k})$, then

$$t^{\surd} = (\bigwedge_{j=1,\ldots,k}(\langle a_j\rangle\top \wedge [a_j]t^{\surd}_{a_j})).$$

We use the formula $t^{\Box\surd}$ to denote that the protocol t is subjectively enabled:

- if $t = (x)$, then $t^{\Box\surd} = \top$.
- if $t = (x, a_1, t_{a_1}) + \ldots + (x, a_k, t_{a_k})$, then

$$t^{\Box\surd} = (\bigwedge_{j=1,\ldots,k}(\Box\langle a_j\rangle\top \wedge \Box[a_j]t^{\Box\surd}_{a_j})).$$

It is straightforward to check that these definitions work as intended:

Lemma 2. *For any protocol T and model $M = (W, \Rightarrow, \rightsquigarrow, V)$, for each $w \in W$, $M, w \models t^{\surd}$ iff enabled(t, w) holds, and $M, w \models t^{\Box\surd}$ iff* s-enabled$((\mathcal{G}, w), t)$ *holds.*

The above reductions from trees to standard PDL formulas suggest that the methods of Kozen and Parikh (1981) to prove completeness of PDL are also applicable in our setting. Our axiomatization follows this "reduction axiom" methodology (i.e., the Segerberg axioms for complex programs) with one important twist: Since the atomic protocols still encode the structure of a tree, we need to provide "reduction axioms" for atomic protocol trees as well. The key idea is to define a formula $push_Q(t, \alpha)$ for $Q \in \{\exists, \forall, \Box\}$ which means that t is (subjectively) enabled and that α holds at all the frontier nodes selected by the relation R^Q_t. These formulas will be defined by induction on the structure of t: For atomic trees $t = (x)$,

(1) $push_{\exists}((x), \alpha) = \alpha$.

(2) $push_{\forall}((x), \alpha) = \alpha$.

(3) $push_{\Box}((x), \alpha) = \Box\alpha$.

For $t = (x, a_1, t_{a_1}) + \ldots + (x, a_k, t_{a_k})$ and $A = \{a_1, \ldots, a_k\}$, we have

(4) $push_{\exists}(t, \alpha) = \bigvee_{a_m \in A} \langle a_m \rangle \langle t_{a_m} \rangle^{\exists} \alpha$.

(5) $push_{\forall}(t, \alpha) = \bigwedge_{a_m \in A} [a_m] \langle t_{a_m} \rangle^{\forall} \alpha$.

(6) $push_{\Box}(t, \alpha) = \bigwedge_{a_m \in A} \Box [a_j] \langle t_{a_j} \rangle^{\Box} \alpha$.

Note that we have not given the corresponding formula for $\langle t \rangle^{\Diamond} \alpha$. This formula is of a different nature than the formulas above. The intended interpretation of $\langle t \rangle^{\Diamond} \alpha$ is that the protocol t is subjectively enabled and α holds at all frontier nodes reached along a *subjective path* in t. Formally, (recall that $Paths(t)$ is the set of maximal paths in T_t), when the path consists of a single node (i.e., $\rho = (x)$) we have:

(1) $cpath((x), \alpha) = \Box \alpha$.

When the path ρ is consists of at least two nodes, we have:

(2) $cpath(\rho, \alpha) = \Box [head(\rho)] cpath(tail(\rho), \alpha)$.

Definition 2.8 (Axiomatization)**.** The **epistemic protocol logic**, denoted EPL, is the smallest set of formulas from $\mathcal{L}_{EPL}$ containing all instances of the following axiom schemes and closed under the following inference rules:

Propositional Tautologies

(1) All instances of propositional tautologies.

Normality Axioms

(2) (a) $\langle \pi \rangle^{\exists} (\alpha_1 \vee \alpha_2) \equiv \langle \pi \rangle^{\exists} \alpha_1 \vee \langle \pi \rangle^{\exists} \alpha_2$

(b) $\Box \alpha_1 \wedge \Box (\alpha_1 \supset \alpha_2) \supset \Box \alpha_2$

Reduction axioms for atomic and composite protocols

(3) $\langle t \rangle^{\forall} \alpha \equiv t^{\surd} \wedge push_{\forall}(t, \alpha)$

(4) $\langle t \rangle^{\exists} \alpha \equiv t^{\surd} \wedge push_{\exists}(t, \alpha)$

(5) $\langle t \rangle^{\Box} \alpha \equiv t^{\Box \surd} \wedge push_{\Box}(t, \alpha)$

(6) $\langle t \rangle^{\Diamond} \alpha \equiv t^{\Box \surd} \wedge \bigvee_{\rho \in Paths(t)} cpath(\rho, \alpha)$

for $Q \in \{\exists, \forall, \Box, \Diamond\}$

(7) $\langle \pi_1 \cup \pi_2 \rangle^{Q} \alpha \equiv \langle \pi_1 \rangle^{Q} \alpha \vee \langle \pi_2 \rangle^{Q} \alpha$

(8) $\langle \pi_1 ; \pi_2 \rangle^{Q} \alpha \equiv \langle \pi_1 \rangle^{Q} \langle \pi_2 \rangle^{Q} \alpha$

(9) $\langle \pi^* \rangle^{Q} \alpha \equiv \alpha \vee \langle \pi \rangle^{Q} \langle \pi^* \rangle^{Q} \alpha$

Inference rules

$$(MP)\ \frac{\alpha,\ \alpha \supset \beta}{\beta} \qquad (NG)\ \frac{\alpha}{[a]\alpha} \qquad (KG)\ \frac{\alpha}{\Box\alpha}$$

$$(IND_Q)\ \frac{\langle\pi\rangle^Q\alpha \supset \alpha}{\langle\pi^*\rangle^Q\alpha \supset \alpha} \quad \text{for } Q \in \{\exists, \forall, \Box, \Diamond\}$$

Some remarks are in order. First, restricting attention to *finite* trees ensures that that the disjunction in axiom A6 is finite. Second, note that normality axioms for $\langle\pi\rangle^\forall$ and $\langle\pi\rangle^\Box$ are *not* valid. Finally, since the action modalities make assertions about the frontier of trees (and forests), the relation R_π^Q is not "upward closed." Nonetheless, the usual PDL axiom for composite programs is still sound:

Proposition 2. *$\langle\pi_1;\pi_2\rangle^Q\alpha \equiv \langle\pi_1\rangle^Q\langle\pi_2\rangle^Q\alpha$ is valid for $Q \in \{\exists, \forall, \Box, \Diamond\}$.*

Proof. We give a proof for the case when $Q = \forall$, the other cases are similar. Suppose that $M, u \models \langle\pi_1;\pi_2\rangle^\forall\alpha$. We will show $M, u \models \langle\pi_1\rangle^\forall\langle\pi_2\rangle^\forall\alpha$. Since $M, u \models \langle\pi_1;\pi_2\rangle^\forall$, there exists $(u, X) \in R^\forall_{\pi_1;\pi_2}$ such that $\forall w \in X$, $M, w \models \alpha$. Hence, there exists $Y = \{v_1, \ldots, v_k\}$ such that $(u, Y) \in R^\forall_{\pi_1}$ and $\forall v_j \in Y$, there exists $X_j \subseteq X$ such that $(v_j, X_j) \in R^\forall_{\pi_2}$ and $\bigcup_{j=1,\ldots,k} X_j = X$. Therefore, $\forall v_k \in Y$, we have $M, v_k \models \langle\pi_2\rangle^\forall\alpha$ and, hence, $M, u \models \langle\pi_1\rangle^\forall\langle\pi_2\rangle^\forall\alpha$.

Conversely, suppose that $M, u \models \langle\pi_1\rangle^\forall\langle\pi_2\rangle^\forall\alpha$. We will show $M, u \models \langle\pi_1;\pi_2\rangle^\forall\alpha$. We have $M, u \models \langle\pi_1\rangle^\forall\langle\pi_2\rangle^\forall\alpha$ iff there exists $(u, Y) \in R^\forall_{\pi_1}$ such that $\forall v_k \in Y$, $M, v_k \models \langle\pi_2\rangle^\forall\alpha$. $M, v_k \models \langle\pi_2\rangle^\forall\alpha$ iff there exists $(v_k, X_k) \in R^\forall_{\pi_2}$ such that $\forall w_k \in X_k$, $M, w_k \models \alpha$. Let $X = \bigcup_k X_k$; from the definition of $R^\forall$ we get $(u, X) \in R^\forall_{\pi_1;\pi_2}$. Hence, $M, u \models \langle\pi_1;\pi_2\rangle^\forall\alpha$. □

We can now state the two main theorems of this section:

Theorem 1. EPL *is sound and weakly complete with respect to the class of all arenas with imperfect information.*

The proof of this theorem can be found in the full version of the paper (Pacuit and Simon 2010).

Corollary 1. *The satisfiability problem for* EPL *is decidable in nondeterministic double exponential time.*[12]

[12] This is an upper bound; the precise lower bound of the satisfiability problem is left open. The

Remark 2.1. Note that the definition of subjectively enabled considers only *single* steps of the $\rightsquigarrow$ relation. One natural generalization here (which we are exploring in a companion paper) is to consider the *transitive closure* of $\rightsquigarrow$ in Definition 2.3. This suggests extending the language with a $\Box^*$ operator, which in turn may open the door to the many axiomatization issues in epistemic temporal languages with common knowledge (cf. van Benthem and Pacuit 2006, for references and a discussion). Also relevant here are the axiomatizations of *products* of PDL and various epistemic and doxastic logics (Schmidt and Tishkovsky 2008).

We can incorporate the properties discussed above. Recall that a formula $\varphi \in \mathcal{L}_{EPL}$ is valid in an arena (with imperfect information) if it is valid in every model based on the arena. First, note that a standard modal *correspondence* argument (cf. Blackburn et al. 2002, Chapter 3) gives us:

Lemma 3. *Let $\mathcal{G}^I = (W, \{\Rightarrow a\}_{a\in\Sigma}, \rightsquigarrow)$ be an arena with imperfect information. Then,*

- *$\mathcal{G}^I$ satisfies no miracles iff $[a]\Box\alpha \supset \Box[a]\alpha$ is valid.*
- *$\mathcal{G}^I$ satisfies success iff $\Diamond\langle a\rangle\top \supset \langle a\rangle\top$ is valid.*
- *$\mathcal{G}^I$ satisfies awareness iff $\langle a\rangle\top \supset \Box\langle a\rangle\top$ is valid.*
- *$\mathcal{G}^I$ satisfies certainty of actions iff $\Diamond\langle a\rangle\top \supset \Box\langle a\rangle\top$ is valid.*

Furthermore, it is not hard to see that adding the axioms in the above Lemma to the axioms in Definition 2.8 leads to a sound and weakly complete axiomatization of the relevant class of models.

3 Actions, abilities and know-how

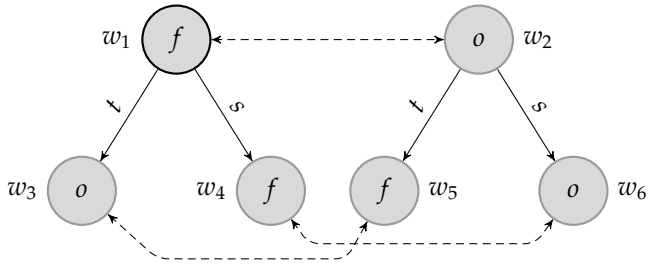
The above discussion focused on the question *under what circumstances can an agent commit to a (joint) protocol or plan, and what can she achieve by doing so*? But, this is only one of many different questions that can be investigated. We mention here one key question:

What is the (formal) difference between an agent knowing that she can achieve φ and knowing how to achieve φ?[13] Much of the work on epistemic extensions of logics

proof is a direct consequence of the proof of the completeness theorem since we construct a finite model.

[13]See (Singh 1999) for a logical analysis of "knowing how" that is related to the framework we develop in this paper.

of actions and abilities has focused on the distinction between *de re* knowledge and *de dicto* knowledge of what agents can achieve (van Benthem 2001, Herzig and Troquard 2006, van der Hoek and Wooldridge 2003b, Jamroga and Agotnes 2007). To illustrate the issue, we use an example from (Herzig and Troquard 2006): Suppose that Ann, who is blind, is standing with her hand on a light switch. She currently does not know whether the light is on or off. The question is does she have the *ability* to turn the light on? Is she *capable* of turning the light on? Does she *know how* to turn the light on? This depends on what we mean by "ability". She has two options available to her: toggle the switch (t) or do nothing (s). This situation is represented by the following arena with imperfect information:



Suppose that the actual state is w_1, so the light is currently off. Now, since Ann is blind, she does not know that the light is off ($w_1 \models \neg\Box f$)[14]. Furthermore, the following formulas are true at w_1: $[t]o$ ("after toggling the light switch (t), the light will be on (o)"), $\neg\Box[t]o$ ("Ann does not know that after toggling the light switch, the light will be on"), $\Box(\langle t\rangle\top \wedge \langle s\rangle\top)$ ("Ann knows that she can toggle the switch (t) and she can do nothing (s)", and $[t]\neg\Box o$ ("after toggling the switch Ann does not know that the light is on"). These formulas describe the basic options available at w_1 and the information Ann has about these options. Consider the basic plan "turn the light on"[15] (denoted by l). Agreeing to this plan commits Ann to a choice between t and o, but this choice can only be made "in the moment" (since, the "correct" option depends on the state of affairs). So, l is a basic protocol consisting of a tree with two branches, one labeled with t and the other labeled with o. We have the following formulas true at state w_1:

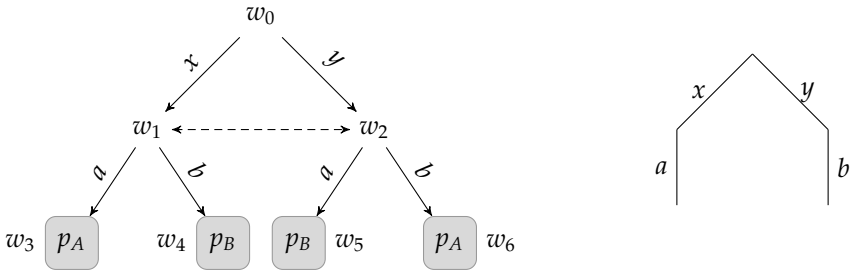
- $\langle l\rangle^{\exists} o \wedge \neg\langle l\rangle^{\forall} o$: executing the plan "turning the light on" can lead to a situation where the light is on, but this is not *guaranteed* (i.e., the plan may fail).

[14] We do not label the modal operator since Ann is the only agent.

[15] Alternatively, we may use the command "make sure the light is on!" to describe this plan.

- $\Box\langle l\rangle^{\exists} o$: Ann knows that she is capable of turning the light on. She has *de re* knowledge that she can turn the light on.
- $\neg\langle l\rangle^{\Diamond} o$: Ann cannot knowingly turn on the light (she does not have *de dicto* knowledge that she can turn the light on): there is no *subjective* path leading to states satisfying o (note that *all* elements of the last element of the subject path must satisfy o).[16]

So, our logical framework can express interesting relationships between a plan π, propositions that can be "brought about" by following π and what the agent(s) knows about π: For example, $\Box\langle\pi\rangle^{\forall}\varphi$ means "the agent *knows that* she can bring about φ by following π", $\langle\pi\rangle^{\forall}\Box\varphi$ means "the agent *can* bring about her knowledge of φ by following π", and $\langle\pi\rangle^{\Box}\varphi$ means "the agent *knows how* to follow π in order to bring about φ". Arguable, the issues discussed above become even more pressing when developing logics of explicit strategies for reasoning about game-theoretic situations van Benthem (2008). In particular, a player may know that she can win the game without actually knowing how (see van Benthem 2001, for a discussion). We conclude this subsection with an initial discussion about how to use our framework for reasoning about strategies in games with imperfect information. Consider an extensive game where Bob moves first (he can choose between x and y) and Ann moves second (she can choose between a and b) without knowledge of Bob's choice:



Suppose that p_A denote a win for Ann and p_B a win for Bob. Let s denote the plan on the right which can be thought of as a strategy for Ann. Indeed, this is a winning strategy for Ann: $\langle s\rangle^{\forall} p_A$ is true at w_0. Furthermore, Ann knows that

[16] It is interesting to note that if t was informative for Ann, so that there is no uncertainty for Ann between states w_3 and w_5, then $\langle l\rangle^{\Diamond} o$ would be true at state w_1. For example, suppose that Ann was not blind, but was standing outside of the room with the door shut and t was the action "open the door".

this is a winning strategy, it is true at w_0 that $\Box\langle s\rangle^{\forall} p_A$ (assume that $w_0 \in \mathcal{I}(w_0)$ for Ann). However, even though this strategy is subjectively enabled for Ann, she does not know how to use this strategy to win the game (in the terminology of van Benthem (2001): the strategy is not *prescriptive*[17]). That is, $\neg\langle s\rangle^{\Box} W$ is true at w_0. These are only some initial observations about how to use our logical systems to reasoning about strategies in imperfect information games — a complete discussion will be left for future work.

4 Conclusions

This paper focuses on the interplay between epistemic reasoning and protocol analysis. In particular, we developed an epistemic protocol logic and discussed what it means for an agent to "subjectively" agree to follow a given protocol. We see this as one step towards addressing the fundamental problem of how to model agents "knowing a protocol, plan or strategy" in situations with imperfect information, and we proved a number of results about our logical system. Besides technical details and proofs, the full version of the paper discusses a number of other issues:

Many agent version: The central issue addressed in this paper is the circumstances under which an agent can "knowingly" agree to follow a protocol or plan. We have seen that even in the single-agent case, this notion is interesting and non-trivial to formalize. However, the situation becomes even more interesting and complex in situations with more than one agent.

Relationship with other logics: There are many other interesting questions to ask about the logical system introduced in the previous section. For example, we can show that $\mathcal{L}_{EPL}$ is strictly more expressive than PDL, but what about concurrent PDL, game logic, the modal μ-calculus, or branching time temporal logic (CTL)? Finding the precise relationship between our epistemic protocol logic and other logical frameworks raises an important question: can we characterize the expressive power of our epistemic protocol language (over the class of arenas with imperfect information). In order to tackle this problem, we need a notion of equivalence between models corresponding to equivalence with respect to $\mathcal{L}_{EPL}$.

[17] This should be contrasted with a strategy that is *uniform*. In our terminology, a protocol π is *uniform* if it is subjectively enabled and it is prescriptive for φ if $\langle\pi\rangle^{\Box}\varphi$ is true at the root node. van Benthem (2001) showed that in games with perfect recall a winning strategy for player i is uniform iff it is prescriptive (for the proposition expressing that player i won the game).

References

P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Campbridge University Press, 2002.

M. Bratman. *Intention, Plans and Practical Reason*. Harvard University Press, 1987.

J. Broersen. A logical analysis of the interaction between 'obligation-to-do' and 'knowingly doing'. In L. van der Torre and R. van der Meyden, editors, *Proceedings 9th International Workshop on Deontic Logic in Computer Science (DEON'08)*, volume 5076 of *Lecture Notes in Computer Science*, pages 140–154. Springer, 2008.

P. R. Cohen and H. Levesque. Intention is choice with committment. *Artificial Intelligence*, 42(3):213 — 261, 1990.

C. B. Cross. 'can' and the logic of ability. *Philosophical Studies*, 50(1):53 – 64, 1986.

D. Elgesem. The modal logic of agency. *Nordic Journal of Philosophical Logic*, 2 (2):1 – 46, 1997.

R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning about Knowledge*. The MIT Press, 1995.

R. Fagin, J. Halpern, Y. Moses, and M. Vardi. Knowledge-based programs. *Distributed Computing*, 10(4):199 – 225, 1997.

G. Governatori and A. Rotolo. On the axiomatization of elgesem's logic of agency and ability. *Journal of Philosophical Logic*, 34(4):403–431, 2005. doi: http://dx.doi.org/10.1007/s10992-004-6368-1.

J. Halpern and Y. Moses. Knowledge and common knowledge in a distributed environment. *Journal of the ACM*, 37(3):549 – 587, 1990.

J. Y. Halpern, R. van der Meyden, and M. Y. Vardi. Complete axiomatizations for reasoning about knowledge and time. *SIAM J. Comput.*, 33(3):674–703, 2004.

A. Herzig and N. Troquard. Knowing how to play: uniform choices in logics of agency. In *AAMAS '06: Proceedings of the fifth international joint conference on autonomous agents and multiagent systems*, pages 209–216, New York, NY, USA,

2006. ACM. ISBN 1-59593-303-4. doi: http://doi.acm.org/10.1145/1160633.1160666.

J. Horty. *Agency and Deontic Logic*. Oxford University Press, 2001.

T. Hoshi. *Epistemic Dynamics and Protocol Information*. PhD thesis, Stanford University, 2009.

T. Icard, E. Pacuit, and Y. Shoham. Joint revision of beliefs and intentions. In *Proceedings of KR 2010*, 2010.

W. Jamroga and T. Agotnes. Constructive knowledge: what agents can achieve under imperfect information. *Journal of Applied Non-Classical Logics*, 17(4):423–475, 2007.

D. Kozen and R. Parikh. An elmentary proof of the completness of PDL. *Theoretical Computer Science*, 14:113 – 118, 1981.

E. Lorini and A. Herzig. A logic of intention and attempt. *Synthese*, 163(1):45 – 77, 2008.

E. Lorini, F. Schwarzentruber, and A. Herzig. Epistemic games in modal logic: joint actions, knowledge and preferences all together. In *Proceedings of LORI'09*, pages 212–226. Springer-Verlag, 2009.

J.-J. Meyer and F. Veltman. Intelligent agents and common sense reasoining. In P. Blackburn, J. van Benthem, and F. Wolter, editors, *Handbook of Modal Logic*, pages 991 – 1029. Elsevier, 2007.

J.-J. Meyer, W. van der Hoek, and B. van Linder. A logical approach to the dynamics of commitments. *Artificial Intelligence*, 113:1 – 40, 1999.

E. Pacuit and S. Simon. Reasoning with protocols under imperfect information. Manuscript, 2010.

R. Parikh and R. Ramanujam. A knowledge based semantics of messages. *Journal of Logic, Language and Information*, 12:453 – 467, 2003.

R. Ramanujam and S. Simon. Dynamic logic of tree composition. In *Perspectives in Concurrency Theory*, pages 408–430. CRC Press, 2009.

R. Schmidt and D. Tishkovsky. On combinations of propositional dynamic logic and doxastic modal logics. *Journal of Logic, Language and Information*, 17 (1):109–129, 2008.

M. P. Singh. Know-how. In M. Wooldridge and A. Rao, editors, *Foundations of Rational Agency*, pages 105 – 132, 1999.

J. van Benthem. Extensive games as process models. *Journal of Logic, Language and Information*, pages 289 – 313, 2001a.

J. van Benthem. Games in dynamic epistemic logic. *Bulletin of Econ. Research*, 53(4):219 – 248, 2001b.

J. van Benthem. In praise of strategies. Technical report, ILLC Technical Reports, 2008.

J. van Benthem. *Logical Dynamics of Information and Interaction*. Cambridge University Press, 2010.

J. van Benthem and E. Pacuit. The tree of knowledge in action: Towards a common perspective. In G. Governatori, I. Hodkinson, and Y. Venema, editors, *Proceedings of Advances in Modal Logic Volume 6*, pages 87 – 106. King's College Press, 2006.

J. van Benthem, J. Gerbrandy, T. Hoshi, and E. Pacuit. Merging frameworks for interaction. *Journal of Philosophical Logic*, 38(5):491–526, 2009.

W. van der Hoek and M. Wooldridge. Towards a logic of rational agency. *Logic Journal of the IGPL*, 11(2):135 – 160, 2003a.

W. van der Hoek and M. Wooldridge. Cooperation, knowledge, and time: Alternating-time temporal epistemic logic and its applications. *Studia Logica*, 75(1):125–157, 2003b.

W. van der Hoek, B. van Linder, and J.-J. Meyer. Formalising abilities and opportunities of agents. *Fundamenta Informaticae*, 34:1 – 49, 1998.

R. van der Meyden. Finite state implementations of knowledge-based programs. In *Proceedings of the Conference on Foundations of Software Technology and Theoretical Computer Science*, volume 1180 of *Lecture Notes in Computer Science*, pages 262 – 273, 1996.

J. van Eijck, L. Kuppusamy, and Y. Wang. Verifying epistemic protocols under common knowledge. In A. Heifetz, editor, *Proceedings of TARK*, pages 257 – 266, 2009.

Y. Wang. *Epistemic Modelling and Protocol Dynamics*. PhD thesis, CWI, 2010.

Toward a Theory of Play: A Logical Perspective on Games and Interaction

Johan van Benthem, Eric Pacuit and Olivier Roy

ILLC Amsterdam & Stanford; TiLPS, Tilburg University; LMU, München
J.vanBenthem@uva.nl, e.j.pacuit@uvt.nl, Olivier.Roy@lrz.uni-muenchen.de

Abstract

Logic and game theory have had a few decades of contacts by now, with the classical results of epistemic game theory as major high-lights. In this paper, we emphasize a recent new perspective toward "logical dynamics", designing logical systems that focus on the actions that change information, preference, and other driving forces of agency. We show how this dynamic turn works out for games, drawing on some recent advances in the literature. Our key examples are the long-term dynamics of information exchange, as well as the much-discussed issue of extensive game rationality. Our paper also proposes a new broader interpretation of what is happening here. The combination of logic and game theory provides a fine-grained perspective on information and interaction dynamics, and we are witnessing the birth of something new which is not just logic, nor just game theory, but rather a *Theory of Play*.

1 Introduction

For many contemporary logicians, games and social interaction are important objects of investigation. *Actions*, *strategies* and *preferences* are central concepts

in computer science and philosophical logic, and their combination raises interesting questions of *definability*, *axiomatization* and *computational complexity* (Bacharach et al. 1997, van der Hoek and Pauly 2006, Bonanno 2002, van Benthem 2001). Epistemic game theory (cf. Brandenburger 2007) has added one more element to this mix, again familiar to logicians: the role of factual and higher-order information. This much is well-understood, and there are excellent sources, that we need not reproduce here, though we will recall a few basics in what follows.

In this paper we will take one step further, assuming that the reader knows the basics of logic and game theory. We are going to take a look at all these components from a dynamic logical perspective, emphasizing actions that make information flow, change beliefs, or modify preferences – in ways to be explained below. For us, understanding social situations as dynamic logical processes where the participants interactively revise their beliefs, change their preferences, and adapt their strategies is a step towards a more finely-structured theory of rational agency. In a simple phrase that sums it up, this joint offspring "in the making" of logic and game theory might be called a *Theory of Play* instead of a theory of games.

The paper starts by laying down the main components of such a theory, a logical take on the dynamics of actions, preferences, and information (Sections 1.1 and 1.2). We then show that this perspective has already shed new light on the long-term dynamics of information exchange, Section 2, as well as on the question of extensive game rationality, Section 3. We conclude with general remarks on the relation between logic and game theory, pleading for cross-fertilization instead of competition. This paper is introductory and programmatic throughout. Our treatment is heavily based on evidence from a number of recent publications demonstrating a variety of new developments.

1.1 An Encounter Between Logic and Games: Extensive Games

A first immediate observation is that games as they stand are natural models for many existing logical languages: epistemic, doxastic and preference logics, as well as conditional logics and temporal logics of action. We do not aim at encyclopedic description of these systems — van der Hoek and Pauly (2006) is a relatively up-to-date overview. This section just gives some examples setting the scene for our later more detailed dynamic-logic analyses.

One rich source of logical structure are the usual strategic games[1]. In this

[1]See the extended version (van Benthem et al. 2011) for a discussion and (de Bruin 2010, Bonanno

contribution, however, we focus on the more fine-structured format of extensive games, showing how it offers a natural meeting point with logic. We will demonstrate this with a case study of *Backwards Induction*, a famous benchmark at the interface, treated in a slightly novel way. Our treatment in this section will be rather classical, that is static and not information-driven. However, in Section 3 we return to the topic, giving it a dynamic, epistemic twist.

Dynamic logic of actions and strategies. The first thing to note is that the sequential structure of players' actions in an extensive game lends itself to logical analysis. A good system to use for this purpose is *propositional dynamic logic* (*PDL*), originally designed to analyze programs and computation (see Harel et al. 2000, for the original motivation and subsequent theory). Let Act be a set of primitive actions. An *action model* is a tuple $\mathcal{M} = \langle W, \{R_a \mid a \in \mathsf{Act}\}, V\rangle$ where W is an abstract set of states, or stages in an extensive game, and for each $a \in \mathsf{Act}$, $R_a \subseteq W \times W$ is a binary *transition relation* describing possible transition from states w to w' via action a. On top of this atomic repertoire, the tree structure of extensive games supports complex action expressions, constructed by the standard regular operations of "indeterministic choice" ($\cup$), "sequential composition" (;) and "unbounded finitary iteration" (*: Kleene star):

$$\alpha \;:=\; a \mid \alpha \cup \beta \mid \alpha;\beta \mid \alpha^*$$

This syntax recursively defines complex relations in action models:

- $R_{\alpha\cup\beta} := R_\alpha \cup R_\beta$
- $R_{\alpha;\beta} := R_\alpha \circ R_\beta$
- $R_{\alpha^*} := \cup_{n\geq 0} R_\alpha^n$. $R_\alpha^0 = Id$ (the identity relation) and $R_\alpha^{n+1} = R_\alpha^n \circ R_\alpha$.

The key dynamic modality $[\alpha]\varphi$ now says that "after the move described by the program expression α is taken, φ is true":

$$\mathcal{M}, w \models [\alpha]\varphi \text{ iff for each } v, \text{ if } wR_\alpha v \text{ then } \mathcal{M}, v \models \varphi$$

PDL has been used for describing solution concepts on extensive games by many authors (cf. Harrenstein et al. 2003, van der Hoek and Pauly 2006, van Benthem 2001). An extended discussion of logics that can explicitly define strategies in extensive games is found in (van Benthem 2008).

2008, van Benthem 2005, Lorini et al. 2009) for further illustrations of logics on strategic games.

Adding preferences: the case of Backwards Induction. As before, a complete logical picture must bring in players' preferences on top of *PDL*, along the lines of our earlier modal preference logic. To show how this works, we consider a key pilot example: the Backwards Induction (*BI*) algorithm. This procedure marks each node of an extensive game tree with *values* for the players (assuming that distinct end nodes have different utility values):[2]

BI **Algorithm**: *At end nodes, players already have their values marked. At further nodes, once all daughters are marked, the player to move gets her maximal value that occurs on a daughter, while the other, non-active player gets his value on that maximal node.*

The resulting strategy for a player selects the successor node with the highest value. The resulting set of moves for all players (still a function on nodes given our assumption on end nodes) is the "*bi* strategy".

Relational strategies and set preference. But to a logician, a strategy is best viewed as a subrelation of the total *move* relation. It is an advice to restrict one's next choice in some way, similar to the more general situation where our *plans* constrain our choices. Mathematically, this links up with the usual way of thinking about programs and procedures in computational logic, in terms of the elegant algebra of relations and its logic *PDL* as defined earlier.

When the above algorithm is modified to a relational setting—we can now drop assumptions about unicity at end-points—we find an interesting new feature: special assumptions about players. For instance, it makes sense to take a minimum value for the passive player at a node over all highest-value moves for the active player. But this is a worst-case assumption: my counter-player does not care about my interests after her own are satisfied. But we might also assume that she does, choosing a maximal value for me among her maximum nodes. This highlights an important feature: *solution methods* are not neutral, they encode significant assumptions about players.

One interesting way of understanding the variety that arises here has to do with the earlier modal preference logic. We might say in general that the driving idea of *Rationality* behind relational *BI* is the following:

I do not play a move when I have another whose outcomes I prefer.

[2] In what follows, we shall mainly work with *finite games*, though current dynamic and temporal logics can also deal with infinite games.

But preferences between moves that can lead to different sets of outcomes call for a notion of "lifting" the given preference on end-points of the game to sets of end-points. As we said before, this is a key topic in preference logic, and here are many options: the game-theoretic rationality behind *BI* has a choice point. One popular version in the logical literature is this:

$$\forall y \in Y \exists x \in X \; x <_i y$$

This says that we choose a move with the highest maximal value that can be achieved. A more demanding notion of preference for a set Y over X in the logical literature (von Wright 1963) is the $\forall\forall$ clause that

$$\forall y \in Y \forall x \in X \; x <_i y$$

Here is what relational *BI* looks like when we follow the latter stipulation, which makes Rationality less demanding, and hence the method more cautious:

> First mark all moves as *active*. Call a move *a dominated* if it has a sibling move all of whose reachable endpoints via active nodes are preferred by the current player to all reachable endpoints via *a* itself. The second version of the *BI* algorithm works in stages:
>
> At each stage, mark dominated moves in the $\forall\forall$ sense of preference as *passive*, leaving all others active.
>
> Here "reachable endpoints" by a move are all those that can be reached via a sequence of moves that are still active at this stage.

We will analyze just this particular algorithm in our logics to follow, but our methods apply much more widely.

Defining Backwards Induction in logic. Many logical definitions for the *BI* strategy have been published (cf. again the survey in van der Hoek and Pauly 2006, Section 3). Here is a modal version combining the logics of action and preferences presented earlier – significantly, involving operator commutations between these:

Theorem 1 (van Benthem et al. (2006))**.** *For each extensive game form, the strategy profile σ is a backward induction solution iff σ is played at the root of a tree satisfying the following modal axiom for all propositions p and players i:*

$$(turn_i \wedge \langle \sigma^* \rangle (end \wedge p)) \rightarrow [move_i] \langle \sigma^* \rangle (end \wedge \langle \geq_i \rangle p)$$

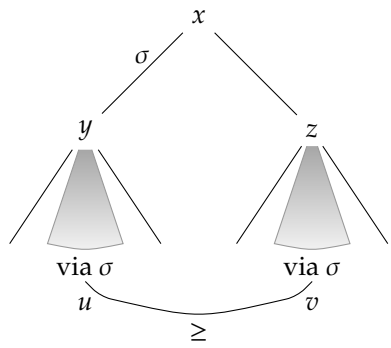
Here $move_i = \bigcup_{a \text{ is an } i\text{-move}} a$, $turn_i$ is a propositional variable saying that it is i's turn to move, and *end* is a propositional variable true at only end nodes. Instead of a proof, we merely develop the logical notions involved a bit further.

The meaning of the crucial axiom follows by a modal frame correspondence (Blackburn et al. 2002, Chapter 3).[3] Our notion of Rationality reappears:

Fact 5. *A game frame makes* $(turn_i \wedge [\sigma^*](end \to p)) \to [move_i]\langle\sigma^*\rangle(end \wedge \langle pref_i\rangle p)$ *true for all i at all nodes iff the frame has this property for all i:*

RAT*: No alternative move for the current player i guarantees outcomes via further play using σ that are all strictly better for i than all outcomes resulting from starting at the current move and then playing σ all the way down the tree.*

A typical picture to keep in mind here, and also later on in this paper, is this:



More formally, *RAT* is this *confluence property* for action and preference:

$$\text{CF} \qquad \bigwedge_i \forall x \forall y((turn_i(x) \wedge x\ \sigma\ y) \to$$

$$(x\ move\ y \wedge \forall z(x\ move\ z \to \exists u \exists v(end(u) \wedge end(v) \wedge y\ \sigma^*\ v \wedge z\ \sigma^*\ u \wedge u \leq_i v)))$$

Now, a simple inductive proof on the depth of finite game trees shows for our cautious algorithm that:

Theorem 2. *BI is the largest subrelation S of the* move *relation in a game with (a) S has a successor at each intermediate node, (b) S satisfies CF.*

This result is not very deep, but it opens a door to a whole area of research.

[3] "Game frames" here are extensive games extended with one more binary relation σ.

The general view: fixed-point logics for game trees. We are now in the realm of a well-known logic of computation, viz. *first-order fixed-point logic LFP(FO)* (Ebbinghaus and Flum 1995). The above analysis really tells us:

Theorem 3. *The BI relation is definable as a greatest-fixed-point formula in the logic LFP(FO).*

Here is the explicit definition in *LFP*(*FO*):

$$BI(x,y) = \nu S.xy \cdot x \; move \; y \wedge \bigwedge_i (Turn_i(x) \rightarrow \forall z(x \; move \; z \rightarrow$$

$$\exists u \exists v(end(u) \wedge end(v) \wedge S.yv \wedge S.zu \wedge u \leq_i v)))$$

The crucial feature making this work is a typical logical point: the occurrences of the relation S in the property CF are *syntactically positive*, and this guarantees upward monotonic behaviour. We will not go into technical details of this connection here, except for noting the following.

Fixed-point formulas in computational logics like this express at the same time static definitions of the *bi* relation, and procedures computing it.[4] Thus, fixed-point logics are an attractive language for extensive games, since they analyze both the statics and dynamics of game solution.

This first analysis of the logic behind extensive games already reveals the fruitfulness of putting together logical and game-theoretical perspectives. But it still leaves untouched the dynamics of deliberation and information flow that determine players' expectations and actual play as a game unfolds, an aspect of game playing that both game theorists and logicians have extensively studied in the last decades. In what follow we make these features explicit, deploying the full potential of the fine-grained Theory of Play that we propose.

1.2 Information Dynamics

The background to the logical systems that follow is a move that has been called a "Dynamic Turn" in logic, making informational acts of inference, but also observations, or questions, into explicit first-class citizens in logical theory

[4]One can use the standard defining sequence for a greatest fixed-point, starting from the total *move* relation, and see that its successive decreasing approximation stages S^k are exactly the 'active move stages' of the above algorithm. This and related connections have been analyzed in greater mathematical detail in (Gheerbrant 2010).

that have their own valid laws that can be brought out in the same mathematical style that has served standard logic so well for so long. The program has been developed in great detail in (van Benthem 1996; 2010) drawing together a wide range of relevant literature, but we will only use some basic components here: single events of information change and, later on in this paper, longer-term interactive processes of information change. The two particular informational events that will be used in this paper are public announcements of hard information, as described by the dynamic-epistemic system **PAL**, and public "radical upgrades" with soft information that change current plausibility relations. In both cases, we are interested in single steps of model change, but also in patterns that emerge in iterated long-term behavior. Towards the end of the paper, we will also briefly refer to other dynamic components of rational agency, with dynamic logics for acts of strategy change, or even preference change. We assume the reader is familiar with this general approach and refer to our extended version (van Benthem et al. 2011) for an introduction and pointers to the relevant literature.

2 Long-term Information Dynamics

We now discuss a first round of applications of the main components of the Theory of Play outlined in the previous sections. We leave aside games for the moment, and concentrate on the dynamic of information in interaction. These applications have in common that they use single update steps, but then iterate them, according to what might be called "protocols" for conversation, learning, or other relevant processes. It is the resulting limit behavior that will mainly occupy us in this section.

We first consider agreement theorems, well known to game theorists, showing how repeated conditioning and public announcements lead to consensus in the limit. This opens the door a general analysis of fixed-points of repeated attitude changes, raising new questions for logic as well as for interactive epistemology. Next we discuss underlying logical issues, including extensions to scenarios of belief merge and formation of group preferences in the limit. Finally we return to a concrete illustration: viz. learning scenarios, a fairly recent chapter in logical dynamics, at the intersection of logic, epistemology, and game theory.

2.1 Agreement Dynamics

Agreement Theorems, introduced in (Aumann 1976), show that common knowledge of disagreement about posterior beliefs is impossible given a common prior. Various generalizations have been given to other informational attitudes, such as probabilistic common belief (Monderer and Samet 1989) and qualitative non-negatively introspective "knowledge" (Samet 2010). These results naturally suggest dynamic scenarios, and indeed Geanakoplos and Polemarchakis (1982) have shown that agreement can be dynamically reached by repeated Bayesian conditioning, given common prior beliefs.

The logical tools introduced above provide a unifying framework for these various generalizations, and allow to extend them to other informational attitudes. For the sake of conciseness, we will not cover static agreement results in this paper. The interested reader can consult (Bonanno and Nehring 1997, Dégremont and Roy 2009).

For a start, we will focus on a comparison between agreements reached via conditioning and via public announcements, reporting the work of (Dégremont and Roy 2009). In the next section, we show how generalized scenarios of this sort can also deal with softer forms of information change, allowing for diversity in update policies within groups.

Repeated Conditioning Lead to Agreements. The following example, inspired by a recent Hollywood production, illustrates how agreements are reached by repeated belief conditioning:

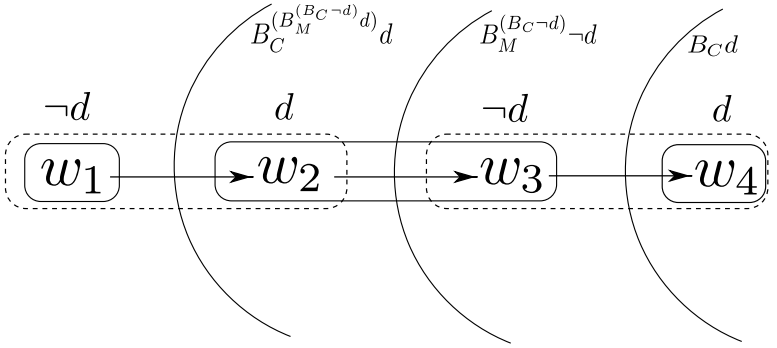


Figure 1: Cobb and Mal on the window ledge.

Example 1. Cobb and Mal are standing on a window ledge, arguing whether they are dreaming or not. Cobb needs to convince Mal, otherwise dreadful consequences will ensue. For the sake of the example, let us assume that Cobb knows they are not dreaming, but Mal mistakenly believes that they are: state w_1 in Figure 1. The solid and dashed rectangles represent, respectively, Cobb's and Mal's hard information. The arrow is their common plausibility ordering.

With some thinking, Mal can come to agree with Cobb. The general procedure for achieving this goes as follows: A *sequence of simultaneous belief conditioning acts* starts with the agents' simple belief about φ, i.e. for all i, the first element $\mathbb{B}_{1,i}$ in the sequence is $B_i\varphi$ if $\mathcal{M}, w \models B_i\varphi$, and $\neg B_i\varphi$ otherwise. Agent i's beliefs about φ at a successor stage are defined by taking her beliefs about φ, conditional upon learning the others' belief about φ at that stage. Formally, for two agents i, j then: $\mathbb{B}_{n+1,i} = B_i^{\mathbb{B}_{n,j}\varphi}\varphi$ if $\mathcal{M}, w \models B_i^{\mathbb{B}_{n,j}\varphi}\varphi$, and $\neg B_i^{\mathbb{B}_{n,j}\varphi}\varphi$ otherwise.[5]

Following the zones marked with an arc in Figure 1, the reader can check that, at w_1, Mal needs three rounds of conditioning to switch her belief about their waking, and thus reach an agreement with Cobb. Her belief stays the same upon learning that Cobb believes that they are not dreaming. Let us call this fact φ. The turning point occurs when she learns that Cobb would not change his mind even if he would learn φ. Conditional on this, she now believes that they are indeed not dreaming. Note that Cobb's beliefs stay unchanged throughout, since he knows the true state at the outset.

Iterated conditioning thus leads to agreement, given common priors. Indeed, conditioning induces a decreasing map from subsets to subsets, which guarantees the existence of a fixed points, where all agent's conditional beliefs stabilize. Once the agents have reached this fixed-point, they have eliminated all higher-order uncertainties concerning the posteriors beliefs about φ of the others. Their posteriors beliefs are now common knowledge:

Theorem 4 (Dégremont and Roy (2009))**.** *At the fixed-point n of a sequence of simultaneous conditioning acts on φ, for all $w \in W$ and $i \in I$, we have that:*

$$\mathcal{M}, w \models C_I(\bigwedge_{i \in I} \mathbb{B}_{n,i}\varphi)$$

The reader accustomed to static agreement theorems will see that we are now only a small step away from concluding that sequences of simultaneous conditionings lead to agreements, as it is indeed the case in our example. Since common prior and common belief of posteriors suffice for agreement, we get:

[5] This definition is meant to fix intuition only. Full details on how to deal with *infinite scenarios*, here and later, are in the cited paper.

Corollary 1. *Take any sequence of conditioning acts for a formula φ, as defined above, in a finite model with common prior. At the fixed point of this sequence, either all agents believe φ or they all don't believe φ.*

This recasts, in our logical framework, the result of (Geanakoplos and Polemarchakis 1982), showing how "dialogs" lead to agreements. Still, belief conditioning has a somewhat private character. In the example above, Cobb remains painfully uncertain of Mal's thinking process until he sees her changing her mind, that is until she makes the last step of conditioning. Luckily for Cobb, they can do better, as we will now proceed to show.

Repeated Public Announcements Lead to Agreements. Figure 2 shows another scenario, where Cobb and Mal publicly and repeatedly announce their beliefs at w_1. They keep announcing the same thing, but each time, this induces important changes in both agents' higher-order information. Mal is led stepwise to realize that they are not dreaming, and crucially, Cobb also knows that Mal receives and processes this information. As the reader can check, at each step in the process, Mal's beliefs are common knowledge.

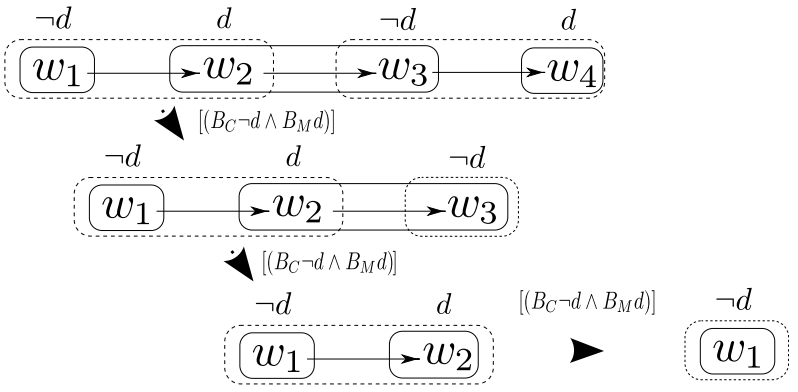


Figure 2: Cobb and Mal's discussion on the window ledge.

One again, Figure 2 exemplifies a general fact. We first define a *dialogue about* φ as a sequence of public announcements. Let $\mathcal{M}, w$ be a finite pointed epistemic-doxastic model.[6] Now let $\mathbb{B}^w_{1,i}$, *i's original belief state at* w, be $B_i\varphi$

[6]Our analysis also applies to infinite models: see the cited papers.

if this formula holds at w, and $\neg B_i\varphi$, otherwise. Agent i's $n+1$ belief state, written $\mathbb{B}^w_{n+1,i}$, is defined as $[\bigwedge_{j\in I}\mathbb{B}^w_{n,j}\varphi]B_i\varphi$ if $\mathcal{M},w \models [\bigwedge_{j\in I}\mathbb{B}^w_{n,j}\varphi]B_i\varphi$, and as $[\bigwedge_{j\in I}\mathbb{B}^w_{n,j}\varphi]\neg B_i\varphi$, otherwise. Intuitively, a dialogue about φ is a process in which all agents in a group publicly and repeatedly announce their posterior beliefs about φ, while updating with the information received in each round.

In dialogues, just like with belief conditioning, iterated public announcements induce decreasing maps between epistemic-doxastic models, and thus are bound to reach a fixed point, where no further discussion is needed. At this point, the protagonists are guaranteed to have reached consensus:

Theorem 5 (Dégremont and Roy (2009))**.** *At the fixed-point $\mathcal{M}_n, w$ of a public dialogue about φ among agents in a group I:*

$$\mathcal{M}_n, w \models C_I(\bigwedge_{i\in I}\mathbb{B}_{n,i})$$

Corollary 2 (Dégremont and Roy (2009))**.** *For any public dialogue about φ, if there is a common prior that is a well-founded plausibility order, then at the fixed-point $\mathcal{M}_n, w$, either all agents believe φ or all do not believe φ.*

As noted in the literature (cf. Geanakoplos and Polemarchakis 1982, Bonanno and Nehring 1997), the preceding dynamics of agreement is one of higher-order information. In the examples above, Mal's information about the ground facts of dreaming or not dreaming, does not change until the very last round of conditioning or public announcement. The information she gets by learning about Cobb's beliefs affects her higher-order beliefs, i.e., what she believes about Cobb's information. This importance of higher-order information flow is a general phenomenon, well-known to epistemic game theorists, which the present logical perspective treats in a unifying way.

Agreements and Dynamics: Further Issues. Here are a few points about the preceding scenarios that invite generalization. Classical agreement results require the agents to be "like-minded" (Bacharach 1985). Our analysis of agreement in dynamic-epistemic logic reveals that this like-mindedness extends beyond the common prior assumption: it also requires the agents to process the information they receive in the same way.[7] One can easily find counter-examples to the agreement theorems when the update rule is not the

[7]Thanks to Alexandru Baltag for pointing out this feature to us.

same for all agents. Indeed, the issue of "agent diversity" is largely unexplored in our logics (but see Liu (2008) for an exception).

A final point is this. While agreement scenarios seem special, to us, they demonstrate a general topic, viz. how different parties in a conversation, say a "Skeptic" and an ordinary person, can modify their positions interactively. In the epistemological literature, this dynamic conversational feature has been neglected – and the above, though solving things in a general way, at least suggests that there might be interesting structure here of epistemological interest.

2.2 Logical Issues about Hard and Soft Limit Behavior

One virtue of our logical perspective is that we can study the above limit phenomena in much greater generality.

Hard information. For a start, for purely logical reasons, iterated public announcement of any formula φ in a model $\mathcal{M}$ must stop at a limit model $lim(\mathcal{M}, \varphi)$ where φ has either become true throughout (it has become common knowledge), or its negation is true throughout.[8] This raises an intriguing open model-theoretic problem of telling, purely from syntactic form, when a given formula is uniformly "self-fulfilling" (the case where common knowledge is reached), or when "self-refuting" (the case where common knowledge is reached of the negation). Game-theoretic assertions of rationality tend to be self-fulfilling, as we shall see in Section 4 below. But there is no stigma attached to the self-refuting case: e.g., the ignorance assertion in the famous Muddy Children puzzle is self-refuting in the limit. Thus, behind our single scenarios, there is a whole area of limit phenomena that have not yet been studied systematically in epistemic logic.[9]

In addition to definability, there is complexity and proof. van Benthem (2001) shows how announcement limit submodels can be defined in various known epistemic fixed-point logics, depending on the syntactic shape of φ. Sometimes the resulting formalisms are decidable, e.g., when the driving assertion φ has "existential positive form", as in the mentioned Muddy Children puzzle, or simple rationality assertions in games.

But these scenarios are still quite special, in that the same assertion gets repeated. There is large variety of further long-term scenarios in the dynamic

[8] We omit some details with pushing the process through infinite ordinals. The final stage is discussed further in terms of "redundant assertions" in (Baltag and Smets 2009a).

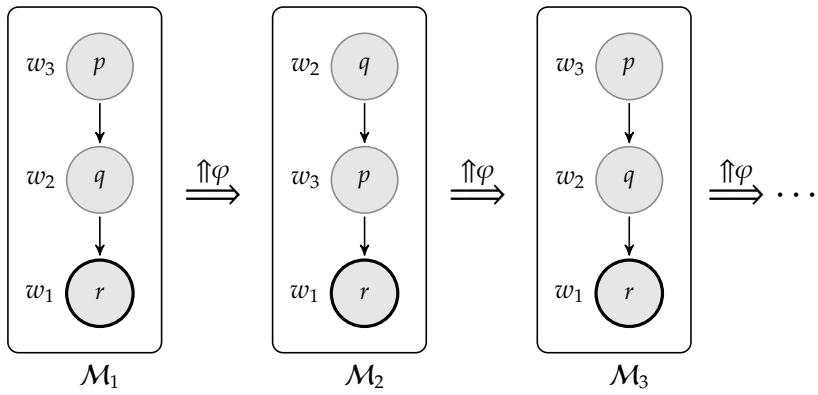
[9] Even in the single-step case, characterizing "self-fulfilling" public announcements has turned out quite involved (cf. Holliday and Icard 2010).

logic literature, starting from the "Tell All" protocols in (van Benthem 2006, Roelofsen 2005, Baltag and Smets 2009b) where agents tell each other all they know at each stage, turning the initial *distributed knowledge* of the group into explicit *common knowledge*.

Soft information. In addition to the limit dynamics of knowledge under hard information, there is the limit behavior of belief, making for more realistic dialog scenarios. This allows for more interesting phenomena in the earlier update sequences. An example is iterated hard information dovetailing agents' opinions, flipping sides in the disagreement until the very last steps of the dialogue (cf. (van Benthem 2010) and (Dégremont 2010, p.110-111)). Such disagreement flips can occur until late in the exchange, but as we saw above, they are bound to stop at some point.

All these phenomena get even more interesting mathematically with dialogs involving soft announcements $[\Uparrow\varphi]$, when limit behavior can be much more complex, as we will see in the next section. Some relevant observations can be found in Baltag and Smets (2009b), and in Section 3 below. First, there need not be convergence at all, the process can oscillate:

Example 2. Suppose that φ is the formula $(r \vee (B^{\neg r}q \wedge p) \vee (B^{\neg r}p \wedge q))$ and consider the one agent epistemic-doxastic models pictured below. Since $[\![\varphi]\!]^{\mathcal{M}_1} = \{w_3, w_1\}$, we have $\mathcal{M}_1^{\Uparrow\varphi} = \mathcal{M}_2$. Furthermore, $[\![\varphi]\!]^{\mathcal{M}_2} = \{w_2, w_1\}$, so $\mathcal{M}_2^{\Uparrow\varphi} = \mathcal{M}_3$. Since, $\mathcal{M}_3$ is the same model as $\mathcal{M}_1$, we have a cycle:



In line with this, players' conditional beliefs may keep changing along the stages of an infinite dialog.[10] But still, there is often convergence at the level of agents' absolute factual beliefs about that the world is like. Indeed, here is a result from Baltag and Smets (2009b):

Theorem 6. *Every iterated sequence of truthful radical upgrades stabilizes all simple non-conditional beliefs in the limit.*

Belief and Preference Merge. Finally, we point at some further aspects of the topics raised here. Integrating agents' orderings through some prescribed process has many similarities with other areas of research. One is *belief merge* where groups of agents try to arrive at a shared group plausibility rder, either as a way of replacing individual orders, or as a way of creating a further group agent that is a most reasonable amalagam of the separate components. And this scenario is again much like those of *social choice theory*, where individual agents have to aggregate preference orders into some optimal public ordering. This naturally involves dynamic analysis of the processes of *delberation* that lead to the eventual act of voting.[11] Thus, the technical issues raised in this section have much wider impact. We may be seeing the contours of a systematic logical study of conversation, deliberation and related social processes.

2.3 Learning

We conclude this section with one concrete setting where many of the earlier themes come together, viz. *formal learning theory*: see (Kelly 1996, Osherson et al. 1986, Schulte 2008). The paradigm we have in mind is identification in the limit of correct hypotheses about the world (cf. Gold (1967) on language learning), though formal learning theory in epistemology has also studied concrete learning algorithms for inquiry of various sorts.

The learning setting shows striking analogies with the dynamic-epistemic logics that we have presented in this paper. What follows is a brief summary of recent work in (Dégremont and Gierasimczuk 2009, Gierasimczuk 2011), to show how our logics link up with learning theory. For broader philosophical backgrounds in epistemology, we refer to (Hendricks 2005). The basic scenario

[10] Infinite iteration of plausibility reordering is in general a non-monotonic process closer to philosophical theories of truth revision in the philosophical literature (cf. Gupta 1982, Herzberger 1982). The technical theory developed on the latter topic in the 1980s may be relevant to our concerns here (cf. Visser 2004).

[11] Van Benthem (2010, Chapter 12), elaborates this connection in more technical detail.

of formal learning theory is one of an agent trying to formulate correct and informative hypotheses about the world, on the basis of an input stream of evidence (in general, an infinite history) whose totality describes what the world is like. At each finite stage of such a sequence, an agent outputs a current hypothesis about the world, which can be modified as new evidence comes in. Success of such a *learning function* in recognition can be of two kinds: either a correct hypothesis is identified uniformly on all histories by some finite stage (the strong notion of "finite identifiability"), or more weakly, each history reaches a point where a correct hypothesis is stated, but when that is may vary according to the history ("identifiability in the limit"). There is a rich mathematical theory of learning functions and what classes of hypotheses can, and cannot, be described by them.

Now, it is not hard to recognize many features here of the logical dynamics that we have discussed. The learning function outputs beliefs, that get revised as new hard information comes in (we think of the observation of the evidence stream as a totally reliable process). Indeed, it is possible to make very precise connections here. We can take the possible hypotheses as our possible worlds, each of which allows those evidence streams (histories of investigation) that satisfy that hypothesis. Then observing successive pieces of evidence is a form of public announcement allowing us to prune the space of worlds. The beliefs involved can be modeled as we did before, by a plausibility ordering on the set of worlds for the agent, which may be modified by successive observations.

On the basis of this simple analogy, Baltag et al. (2010) prove results like the following, making connections very tight:

Theorem 7. *Public announcement-style eliminative update is a universal method: for any learning function, there exists a plausibility order that encodes the successive learning states as current beliefs. The same is true, taking observations as events of soft information, for radical upgrade of plausibility orders.*

Theorem 8. *When evidence streams may contain a finite amount of errors, public announcement-style update is no longer a universal learning mechanisms, but radical upgrade still is.*

With these bridges in place, one can also introduce logical languages in the learning-theoretic universe. Dégremont and Gierasimczuk (2009) show how many notions in learning theory then become expressible in dynamic-epistemic or epistemic-temporal languages, say convergence in the limit as necessary future truth of knowledge of a correct hypothesis about the world. [12] Thus, we

[12]The logical perspective can actually define many further refinements of learning desiderata,

seem to be witnessing the beginning of merges between dynamic logic, belief revision theory and learning theory.

Such combinations of dynamic epistemic logic and learning theory also invite comparison with game theory. Learning, for instance, to coordinate on a Nash equilibrium in repeated games, has been extensively studied, with many positive and negative results—see, for example, (Kalai and Lehrer 1993).[13]

This concludes our exploration of long-term information dynamics in our logical setting. We have definitely not exhausted all possible connections, but we hope to have shown how a general Theory of Play fits in naturally with many different areas, providing a common language between them.

3 Solution Dynamics on Extensive Games

We now return to game theory proper, and bring our dynamic logic perspective to bear on an earlier benchmark example: Backwards Induction. This topic has been well-discussed already by eminent authors, but we hope to add a number of new twists suggesting broader ramifications in the study of agency.

In the light of logical dynamics, the main interest of a solution concept is not its "outcome", its set of strategy profiles, but rather its "process", the way in which these outcomes are reached. Rationality seems largely a feature of procedures we follow, and our dynamic logics are well-suited to focus on that.

3.1 Belief and Soft Plausibility Upgrade

Many foundational studies in game theory view Rationality as choosing a best action *given what one believes* about the current and future behaviour of the players. An appealing alternative take on the *BI* procedure does not eliminate any nodes of the initial game, but rather endows it with "progressive expectations" on how the game will proceeed. This is the plausibility dynamics that we studied in Section 3, now performing a *soft announcement* of **rat**, where the appropriate action is the "radical upgrade" studied earlier. The essential information produced by the algorithm is then in the binary plausibility relations that it creates inductively for players among end nodes in the game,

such as reaching future stages when the agent's knowledge becomes introspective, or when her belief becomes correct, or known.

[13] Many of these results live in a probabilistic setting, but dynamic logic and probability is another natural connection, that we have to forego in this paper.

standing for complete histories or "worlds":

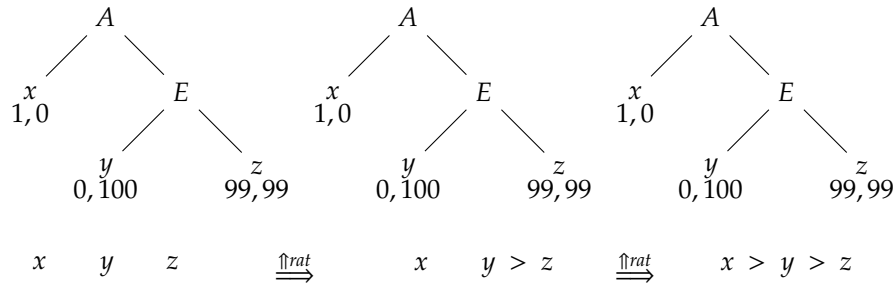
Example [The *BI* outcome in a soft light.] A soft scenario does not remove nodes but modifies the plausibility relation. To implement this, we start with all endpoints of the game tree incomparable.[14] Next, at each stage, we compare sibling nodes, using this notion:

> A move x for player i dominates its sibling y *in beliefs* if the *most plausible* end nodes reachable after x along any path in the whole game tree are all better for the active player than all the *most plausible* end nodes reachable in the game after y.

Rationality* (*rat*∗) says no player plays a move that is dominated in beliefs. Now we perform essentially a radical upgrade $\Uparrow rat^*$:[15]

> If game node x dominates node y in beliefs, make all end nodes reachable from x more plausible than those reachable from y, keeping the old order inside these zones.

This changes the plausibility order, and hence the pattern of dominance-in-belief, so that iteration makes sense. Here are the stages in our earlier example, where letters x, y, z stand for the end nodes of the game:



In the first game tree, going right is not yet dominated in beliefs for A by going left. *rat*∗ only has bite at E's turn, and an upgrade takes place that makes $(0, 100)$ more plausible than $(99, 99)$. After this upgrade, however, going right has now become dominated in beliefs, and a new upgrade takes place, making A's going left most plausible. Here is the general result (cf. van Benthem 2010, van Benthem and Gheerbrant 2010):

[14] Other versions of our scenario would rather make them equi-plausible.

[15] We refer to van Benthem and Gheerbrant (2010) for technical details.

Theorem 9. *On finite trees, the Backwards Induction strategy is encoded in the plausibility order for end nodes created by iterated radical upgrade with rationality-in-belief.*

Again this is "self-fulfilling": at the end of the procedure, the players have acquired common belief in rationality. An illuminating way of proving this uses an idea from Baltag et al. (2009):

Strategies as plausibility relations. Each sub-relation R of the total *move* relation induces a total plausibility order $ord(R)$ on endpoints of a game:

> $x\ ord(R)\ y$ iff, looking up at the first node z where the histories of x, y diverged, if x was reached via an R move from z, then so is y.

More generally, relational strategies correspond one-to-one with "move-compatible" total orders of endpoints. In particular, conversely, each such order $\leq$ induces a strategy $rel(\leq)$. Now we can relate the computation in our upgrade scenario for belief and plausibility to the earlier relational algorithm for BI in Section 1:

Fact 6. *For any game tree $\mathcal{M}$ and any k, $rel((\Uparrow rat^*)^k, \mathcal{M})) = BI^k$.*

Thus, the algorithmic view of Backwards Induction and its procedural doxastic analysis in terms of forming beliefs amount to the same thing. Still, as with our iterated announcement scenario, the dynamic logical view has interesting features of its own. One is that it yields fine-structure to the plausibility relations among worlds that are usually taken as primitive in doxastic logic. Thus games provide an underpinning for the possible worlds semantics of belief that seems of interest per se.

3.2 Logical Dynamic Foundations of Game Theory

We have seen, if only briefly, how dynamic approaches to Backwards Induction focus on the procedure itself as the locus of rationality. Moreover, in doing so, extensionally equivalent definitions can still have interesting intensional differences. For instance, the above analysis of strategy creation and plausibility change seems the most realistic description of the "entanglement" of belief and rational action in the behaviour of agents. But as we will discuss soon, a technical view in terms of fixed-point logics may be the best mathematical approach linking up with other areas.

No matter how we construe them, one key feature of our dynamic announcement and upgrade scenarios is this. Unlike the usual epistemic foundation results, common knowledge or belief of rationality is not assumed, but *produced* by the logic. This reflects our general view that rationality is primarily a property of procedures of deliberation or other logical activities, and only secondarily a property of outcomes of such procedures.

In particular, it has been suggested that game solution may be viewed as a process of a priori "deliberation" by players who are trying to simplify the game through a sequence of considerations. One way of casting this is in terms of iterated public announcements, like we did in the preceding section. The driving assertion this time is an appropriate form of "rationality" as never playing dominated moves, and it can be shown that in the limit, the actual Backwards Induction path is obtained in this way (cf. van Benthem 2007). But in this paper, we take another, more sophisticated road, showing how the deliberation procedure may also be cast as one of creating successive sharper beliefs about how the game will proceed.

3.3 Logics of Game Solutions: General Issues

Our analysis does not just restate existing game-theoretic results, it also raises new issues in the logic of rational agency. Technically, all that has been said in Sections 2 an 3 can be formulated in terms of existing *fixed-point logics* of computation, such as the modal "μ-calculus" and the first-order fixed-point logic *LFP*(*FO*). This link with a well-developed area of computational logic is attractive, since many results are known there, and we may use them to investigate game solution procedures that are quite different from Backwards Induction.[16] But the analysis of game solutions also brings some new logical issues to this area.

Game solution and fragments of fixed-point logics. Game solution procedures need not use the full power of fixed-point languages for recursive procedures. It makes sense to use small decidable fragments where appropriate. Still, it is not quite clear right now what the best fragments are. In particular, our earlier analysis intertwines two different relations on trees: the *move* relation of action and computation, and the *preference* relations for players on endpoints. And the question is what happens to known properties of computational logics when we add such preference relations:

[16] See the dissertation (Gheerbrant 2010) for details, linking up with computational logic.

The complexity of rationality. In combined logics of action and knowledge, it is well-known that apparently harmless assumptions such as Perfect Recall for agents make the validities undecidable, or non-axiomatizable, sometimes even Π^1_1-complete (Halpern and Vardi 1989). The reason is that these assumptions generate commuting diagrams for actions *move* and epistemic uncertainty $\sim$ satisfying a "confluence property"

$$\forall x \forall y((x \; move \; y \wedge y \sim z) \rightarrow \exists u(x \sim u \wedge u \; move \; z))$$

These patterns serve as the basic grid cells in encodings of complex "tiling problems" in the logic.[17] Thus, the logical theory of games for players with perfect memory is more complex than that of forgetful agents (cf. Halpern and Vardi 1989, van Benthem and Pacuit 2006). But now consider the non-epistemic property of rationality studied above, that mixes action and preference. Our key property *CF* in Section 1 had a confluence flavour, too, with a diagram involving action and preference:

$$\forall x \forall y((Turn_i(x) \wedge x \; \sigma \; y) \rightarrow \forall z(z \; move \; z \rightarrow \forall u((end(u) \wedge y \; \sigma^* \; u)$$

$$\rightarrow \exists v(end(v) \wedge z \; \sigma^* v \wedge v \leq_i u))))$$

So, what is the complexity of fixed-point logics for players with this kind of regular behaviour? Can it be that Rationality, a property meant to make behaviour simple and predictable, actually makes its theory complex?

Zooming in and zooming out: modal logics of best action. The main trend in our analysis has been toward making dynamics explicit in richer logics than the usual epistemic-doxastic-preferential ones, in line with the program in van Benthem (2010). But in logical analysis, there are always two opposite directions intertwined: getting at important reasoning patterns by making things more explicit, or rather, by making things less explicit!

In particular, in practical reasoning, we are often only interested in what are our *best actions* without all details of their justification. As a mathematical abstraction, it would then be good to extract a simple surface logic for reasoning with best actions, while hiding most of the machinery:

> Can we axiomatize the modal logic of finite game trees with a *move* relation and its transitive closure, turns and preference relations for players, and a new relation *best* computed by Backwards Induction?

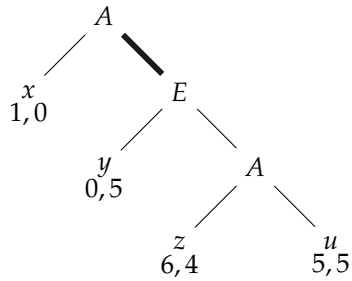
[17] Recall our earlier remarks in Section 1.1 on the complexity of strategic games.

Further logical issues in our framework concern extensions to *infinite games*, games with *imperfect information*, and scenarios with *diverse agents*. See Liu (2008), Dégremont (2010), Zvesper (2010) for some first explorations.

3.4 From Games to their Players

We end by high-lighting a perhaps debatable assumption of our analysis so far. It has been claimed that the very Backwards Induction reasoning that ran so smoothly in our presentation, is incoherent when we try to "replay" it in the opposite order, when a game is actually played.[18]

Example [The 'Paradox of Backwards Induction'.] Recall the style of reasoning toward a Backward Induction solution, as in our earlier simple scenario:



Backwards Induction tells us that A will go left at the start, on the basis of logical reasoning that is available to both players. But then, if A plays *right* (as marked by the thick black line) what should E conclude? Does not this mean that A is not following the *BI* reasoning, and hence that all bets are off as to what he will do later on in the game? It seems that the very basis for the computations in our earlier sections collapses.[19]

Responses to this difficulty vary. Many game-theorists seem under-impressed. The characterization result of Aumann (1995) assumes that players know that rationality prevails throughout.[20] One can defend this behaviour by assuming that the other player only makes isolated mistakes. Baltag, Smets

[18]There is a large literature focused on this "paradox" of backwards induction which we do not discuss here. See, for example, Bicchieri (1989).

[19]The drama is clearer in longer games, when A has many comebacks toward the right.

[20]Samet Samet (1996) calls this "rationality no matter what", a stubborn unshakable belief that players will act rationally later on, even if they have never done so up until now.

and Zvesper Baltag et al. (2009) essentially take the same tack, deriving the *BI* strategy from an assumption of "stable true belief" in rationality, a gentler form of stubbornness stated in terms of dynamic-epistemic logic.

Players' revision policies. We are more inclined toward the line of Stalnaker (1996; 2001). A richer analysis should add an account of the *types of agent* that play the game. In particular, we need to represent the *belief revision policies* of the players, that determine what they will do when making a surprising observation contradicting their beliefs in the course of a game. There are many different options for such policies in the above example, such as "It was just an error, and A will go back to being rational", "A is telling me that he wants me to go right, and I will be rewarded for that", "A is an automaton with a general rightward tendency", and so on.[21] Our analysis so far has omitted this type of information about players of the game, since our algorithms made implicit uniform assumptions about their prior deliberation, as well as what they are going to do as the game proceeds.

This matching up of two directions of thought: *backwards* in "off-line dynamics" of deliberation, and *forwards* in "on-line dynamics" of playing the actual game, is a major issue in its own right, beyond specific scenarios. Belief revision policies and other features of players must come in as explicit components of the theory, in order to deal with the dynamics of how players update knowledge and revise beliefs as a game proceeds.

But all this is exactly what the logical dynamics of Section 1.2 is about. Our earlier discussion has shown how acts of information change and belief revision can enter logic in a systematic manner. Thus, once more, the richer setting that we need for a truly general theory of game solution is a perfect illustration for the general Theory of Play that we have advocated.

4 Conclusion

Logic and game theory form a natural match, since the structures of game theory are very close to being models of the sort that logicians typically study. Our first illustrations reviewed existing work on static logics of game structure, drawing attention to the fixed-point logic character of game solution methods.

[21] One reaction to these surprise events might even be a switch to an entirely new style of reasoning about the game. That would require a more finely-grained *syntax-based* views of revision: cf. the discussion in Velázquez-Quesada (2011).

This suggests a broader potential for joining forces between game theory and computational logic, going beyond specific scenarios toward more general theory. To make this more concrete, we then presented the recent program of "logical dynamics" for information-driven agency, and showed how it throws new light on basic issues studied in game theory, such as agreement scenarios and game solution concepts.

What we expect from this contact is not the solution of problems afflicting game theory through logic, or vice versa, remedying the aches and pains of logic through game theory. Of course, game theorists may be led to new thoughts by seeing how a logician treats (or mistreats) their topics, and also, as we have shown, logicians may see interesting new open problems through the lense of game theory.

But fruitful human relations are usually not therapeutic: they lead to new facts, in the form of shared offspring. In particular, one broad trend behind much of what we have discussed here is this. Through the fine-structure offered by logic, we can see the dynamics of games as played in much more detail, making them part of a general analysis of agency that also occurs in many other areas, from "multi-agent systems" in computer science to social epistemology and the philosophy of action. It is our expectation that the offspring of this contact might be something new, neither fully logic nor game theory: a *Theory of Play*, rather than just a theory of games.

References

R. Aumann. Agreeing to disagree. *Annals of Statistics*, 4:1236 — 1239, 1976.

R. Aumann. Backward induction and common knowledge of rationality. *Games and Economic Behavior*, 8:6 – 19, 1995.

M. Bacharach. Some extensions of a claim of Aumann in an axiomatic model of knowledge. *Journal of Economic Theory*, 37(1):167–190, October 1985.

M. Bacharach, L. Gerard-Varet, P. Mongin, and H. Shin, editors. *Epsitemic Logic and the Theory of Games and Decisions*. Kluwer Academic Publishers, 1997.

A. Baltag and S. Smets. Talking your way into agreement: Belief merge by persuasive communication. In *Proceedings of the Second Multi- Agent Logics, Languages, and Organisations Federated Workshops*, volume 494, pages 129 – 141, 2009a.

A. Baltag and S. Smets. Group belief dynamics under iterated revision: Fixed points and cycles of joint upgrades. In *Proceedings of Theoretical Aspects of Rationality and Knowledge*, 2009b.

A. Baltag, S. Smets, and J. Zvesper. Keep 'hoping' for rationality: a solution to the backward induction paradox. *Synthese*, 169(2):301–333, 2009. URL `http://dx.doi.org/10.1007/s11229-009-9559-z`.

A. Baltag, N. Gierasimczuk, and S. Smets. Truth-tracking by belief revision. Unpublished manuscript, 2010.

C. Bicchieri. Self-refuting theories of strategic interaction: a paradox of common knowledge. *Erkenntnis*, 30:69–85, 1989.

P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Campbridge University Press, 2002.

G. Bonanno. Modal logic and game theory: two alternative approaches. *Risk, Decision and Policy*, 7(3):309–324, 2002.

G. Bonanno. A syntactic approach to rationality in games with ordinal payoffs in: (eds.),. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *Logic and the Foundations of Game and Decision Theory*, Texts in Logic and Games Series, pages 59 – 86. Amsterdam University Press, 2008.

G. Bonanno and K. Nehring. Agreeing to disagree: a survey. Document prepared of an invited lecture at the Workshop on Bounded Rationality and Economic Modelling, July 1997. URL `http://www.econ.ucdavis.edu/faculty/bonanno/wpapers.htm`.

A. Brandenburger. The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory*, 35:465–492, 2007.

B. de Bruin. *Explaining Games: The Epistemic Programme in Game Theory*. Springer, 2010.

C. Dégremont. *The Temporal Mind. Observations on the Logic of Belief Change in Interactive Systems*. PhD thesis, ILLC University of Amsterdam Dissertation Series DS-2010-03, 2010.

C. Dégremont and N. Gierasimczuk. Can doxastic agents learn? on the temporal structure of learning. In X. He, J. Horty, and E. Pacuit, editors, *Proceedings of LORI-II*, pages 90 – 104, 2009.

C. Dégremont and O. Roy. Agreement theorems in dynamic-epistemic logic. In A. Heifetz, editor, *Proceedings of TARK'09 (Theoretical Aspects of Rationality and Knowledge)*. ACM Digital Library, July 2009.

H.-D. Ebbinghaus and J. Flum. *Finite Model Theory*. Springer, 1995.

J. Geanakoplos and H. Polemarchakis. We can't disagree forever. *Journal of Economic Theory*, 28(1):192 – 200, 1982.

A. Gheerbrant. *Fixed-Point Logics on Trees*. PhD thesis, Institute for Logic, Language and Information Dissertation Series DS-2010-08, 2010.

N. Gierasimczuk. *Knowing Ones Limits: Logical Analysis of Inductive Inference*. PhD thesis, ILLC University of Amsterdam, 2011.

E. M. Gold. Language identification in the limit. *Information and Control*, 10 (5):447–474, 1967. URL `http://www.isrl.uiuc.edu/~amag/langev/paper/gold67limit.html`.

A. Gupta. Truth and paradox. *Journal of Philosophical Logic*, 11:1 – 60, 1982.

J. Halpern and M. Vardi. The complexity of reasoning about knowledge and time. *Journal of Computer and System Sciences*, 38:195 – 237, 1989.

D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic*. The MIT Press, 2000.

B. Harrenstein, W. van der Hoek, J.-J. Meyer, and C. Witteveen. A modal characterization of nash equilibrium. *Fundamenta Informaticae*, 57(2-4):281 – 321, 2003.

V. Hendricks. *Mainstream and Formal Epistemology*. Cambridge University Press, 2005.

H. Herzberger. Notes on naive semantics. *Journal of Philosophical Logic*, 11:61 – 102, 1982.

W. Holliday and T. Icard. Moorean phenomena in epistemic logic. In L. Beklemishev, V. Goranko, and V. Shehtman, editors, *Advances in Modal Logic*, volume 8pp., College Publications., pages 178 – 199, 2010.

E. Kalai and E. Lehrer. Rational learning leads to nash equilibrium. *Econometrica*, 61(5):1019 – 1045, 1993.

K. Kelly. *The Logic of Reliable Inquiry*. Oxford University Press, 1996.

F. Liu. *Changing for the Better: Preference Dynamics and Agent Diversity*. PhD thesis, Institute for Logic, Language and Computation (DS-2008-02), 2008.

E. Lorini, F. Schwarzentruber, and A. Herzig. Epistemic games in modal logic: Joint actions, knowledge and preferences all together. In X. He, J. Horty, and E. Pacuit, editors, *LORI-II Workshop on Logic, Rationality and Interaction, Chongqing, China*, pages 212–226. Springer-Verlag, 2009.

D. Monderer and D. Samet. Approximating common knowledge with common beliefs. *Games and Economic Behavior*, 1:170 – 190, 1989.

D. Osherson, M. Stob, and S. Weinstein. *Systems that Learn*. The MIT Press, 1986.

F. Roelofsen. Exploring logical perspectives on distributed information and its dynamics. Master's thesis, ILLC University of Amsterdam (LDC 2005-05), 2005.

D. Samet. Hypothetical knowledge and games with perfect information. *Games and Economic Behavior*, 17(2):230 – 251, 1996.

D. Samet. Agreeing to disagree: The non-probabilistic case. *Games and Economic Behavior*, 69:169 – 174, 2010.

O. Schulte. Formal learning theory. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2008 edition, 2008.

R. Stalnaker. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, 12(02):133 – 163, 1996.

R. Stalnaker. Substantive rationality and backward induction. *Games and Economic Behavior*, 37(2):425 – 435, 2001.

J. van Benthem. *Exploring Logical Dynamics*. CSLI Press, 1996.

J. van Benthem. Games in dynamic epistemic logic. *Bulletin of Econ. Research*, 53(4):219 – 248, 2001.

J. van Benthem. Open problems in logic and games. In S. Artemov, H. Barringer, A. d'Avila Garcez, L. Lamb, and J. Woods, editors, *We Will Show Them! Essays in Honour of Dov Gabbay*, volume 1. College Publications, 2005.

J. van Benthem. One is a lonely number: on the logic of communication. In Z. Chatzidakis, P. Koepke, and W. Pohlers, editors, *Logic Colloquium '02*, volume 27 of *Lecture Notes in Logic*, pages 96 – 129. ASL & A.K. Peters, 2006.

J. van Benthem. Rational dynamics and epistemic logic in games. *International Journal of Game Theory Review*, 9(1):13 – 45, 2007.

J. van Benthem. In praise of strategies. Technical report, ILLC Technical Reports, 2008.

J. van Benthem. *Logical Dynamics of Information and Interaction*. Cambridge University Press, 2010.

J. van Benthem and A. Gheerbrant. Game solution, epistemic dynamics and fixed-point logics. *Fundamenta Informaticae*, 100:1 – 23, 2010.

J. van Benthem and E. Pacuit. The tree of knowledge in action: Towards a common perspective. In G. Governatori, I. Hodkinson, and Y. Venema, editors, *Proceedings of Advances in Modal Logic Volume 6*, pages 87 – 106. King's College Press, 2006.

J. van Benthem, S. van Otterloo, and O. Roy. Preference logic, conditionals, and solution concepts in games. In H. Lagerlund, S. Lindström, and R. Sliwinski, editors, *Modality Matters: Twenty-Five Essays in Honour of Krister Segerberg*, volume 53 of *Uppsala Philosophical Studies*. University of Uppsala, 2006.

J. van Benthem, E. Pacuit, and O. Roy. Toward a theory of play: A logical perspective on games and interaction. *Games*, 2(1):52 – 86, 2011.

W. van der Hoek and M. Pauly. Modal logic for games and information. In P. Blackburn, J. van Benthem, and F. Wolter, editors, *Handbook of Modal Logic*, volume 3 of *Studies in Logic*, pages 1077 – 1148. Elsevier, 2006.

F. R. Velázquez-Quesada. *Small Steps in the Dynamics of Information*. PhD thesis, ILLC University of Amsterdam, 2011.

A. Visser. Semantics and the liar paradox. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 11. Springer, 2004.

G. H. von Wright. *The Logic of Preference*. Edinburgh, 1963.

J. Zvesper. *Playing with Information*. PhD thesis, ILLC University of Amsterdam Dissertation Series DS-2010-02, 2010.

A Logic for Extensional Protocols

Ben Rodenhäuser

University of Groningen
b.rodenhaeuser@gmail.com

Abstract

We study a logic for reasoning about agents that pass messages according to a protocol. Protocols are specified extensionally, as sets of sequences of "legal" actions assigned to each state in a Kripke model. Message-passing events that are licensed by the protocol lead to updates in the style of dynamic epistemic logic. We also consider changes to the protocol by introducing a notion of protocol extension. Using simple scenarios, we demonstrate that the logic is capable of abstractly modeling agents that agree on and carry out plans. While in our general framework, messages are arbitrary objects, we also consider the case that the messages that are passed are sentences of the object language itself. We present three complete calculi, axiomatizing the logic of protocol-based message-passing, the logic of protocol-based message-passing using sentences of epistemic logic and the logic of protocol extension.

Introduction

Rohit Parikh and Ram Ramanujam begin their influential paper "A knowledge-based semantics of messages" (Parikh and Ramanujam (2003), henceforth cited as "Parikh and Ramanujam") by quoting the following passage from a poem by Longfellow:

> *He said to his friend, "if the British march*

by land or sea from the town tonight,
Hang a lantern aloft in the belfry arch
Of the North church tower as a signal light, –
One, if by land, and two, if by sea;
And I on the opposite shore will be, ..."

The authors go on to comment:

> "Here, Paul Revere is setting up a protocol with his friend whereby a signal with two possible values can be used to indicate one of two alternatives. While the hanging of a lantern merely shows a light, doing this at a specific time, in a specific state of the world, by one person, carries a meaning to another who sees the light, *when the two have agreed on a protocol for signaling, and the latter trusts the former to follow the protocol.*" (Parikh and Ramanujam, emphasis in the original).

This paper essentially elaborates on these remarks and analyzes *protocol-based message-passing*. As suggested by the example of Paul Revere, we are interested in two types of events that appear in this context: *message-passing events* that are licensed by a given protocol (e.g., the hanging of a lantern, once a protocol for signaling has been agreed upon); and *protocol extensions* that change the current protocol by allowing new messages to be passed (e.g., the instruction to light a lantern if the British march by land, and two lanterns if they march by sea). Our modeling will remain at a very abstract level; even so, if we think of the events as actions of our agents, our logic can be seen as modeling agents that *commit to plans (or inform each other about their intentions) and act based on these plans*. More generally, since the protocols we consider consist of sequences of messages that convey semantic information to our agents (rather than changing "ground facts" in the world), we are dealing with a logic for reasoning about *epistemic change*.

The work of the paper draws on two research traditions in epistemic logic that offer different perspectives on epistemic change: epistemic temporal logic (ETL) and dynamic epistemic logic (DEL). Authors working in the ETL tradition (cf. Fagin et al. (1995) and Parikh and Ramanujam) have adopted, essentially, a *global perspective* on epistemic change: one views all possible ways in which epistemic states might evolve in their totality, spelling out everything that can happen to the knowledge of the agents. This view has also been called the "grand stage" perspective (van Benthem 2010). Conceptually, it is closely related to the way in which an "extensive form" spells out a game: one draws

a tree representing all possible sequences of "legal moves" (Osborne and Rubinstein 1994). The Kripke models used in ETL, then, contain, besides *epistemic* accessibility relations, a second family of relations capturing *temporal* transitions from one point in time to another. In this way, ETL models capture protocols: each path through such a model represents a sequence of events that are "legal" according to, or "licensed" by the protocol.

Researchers working in the second, more recent DEL tradition (Plaza 1989, Gerbrandy 1999, Baltag et al. 1999, van Benthem 2010) have adopted a *local perspective* on epistemic change: one zooms in on a specific "system state" and studies how the system may be affected by various types of events that change the information the agents have about the world and each other. The models used thus merely represent the epistemic state of the agents at a given, specific point in time. Instead of adding temporal transition relations to such models, new epistemic states are computed from old ones using model transformations, or "epistemic updates". Rather than making a transition from a state *in a model* to another, "later" state *in the same model* (as in ETL formalisms), one considers transitions *from models to models*. Because of this feature, DEL systems can be said to provide an "update semantics".

It has been observed that the standard DEL approach is not directly useable for the study of protocols (for a discussion with further references, see Hoshi (2010)). In terms of our above game analogy: in DEL, one zooms in on a certain "game-playing situation" that might occur while playing a game. But there is no direct formal analogue to the protocol that governs which actions constitute legal moves and which do not (see, however, Baltag (2002) for a "rule-based" DEL approach to games). In public announcement logic (PAL), e.g., an assumption that is hardwired into the setting is that *every true sentence can be announced*; there is no immediate way to impose further constraints on possible announcements. But such constraints are precisely what the temporal transition relations used in ETL capture. So standard DEL does not offer a notion of *protocol-based update*, where "protocol-based update" is a shorthand for *an update induced by a message that is interpreted according to a given protocol*.

The issue of how to combine the DEL perspective on epistemic change with a useful notion of protocols is thus important—not least because protocols have long been seen as important in applications of epistemic logic to multi-agent systems, game theory and distributed computing (Lamport et al. 1982, Ladner and Reif 1986, Halpern and Moses 1990). While there is a growing logical literature on various aspects of protocols (cf. Wang (2010) and Pacuit and Simon (2010) for discussion and further references), we are specifically interested in addressing this issue in this paper.

A recent line of work has studied ways to "merge" the DEL and the ETL perspective on epistemic change (van Benthem et al. 2009, Hoshi 2009, Dégremont 2010). This research has led to new links between DEL and ETL in terms of characterization theorems, and to the identification of new axiomatic theories such as the logic TPAL, temporal public announcement logic (van Benthem et al. (2009), henceforth cited as "van Benthem *et al.*"). Semantically, van Benthem *et al.* essentially adopt the afore-mentioned *global* perspective on epistemic change: the authors are working with a standard ETL semantics restricted to particular classes of temporal models. In the case of TPAL, these temporal models are the *ETL models generated by PAL-protocols*. The truth predicate of TPAL, however, is not stated in terms of model transformations. If one regards—as is usually done, cf. Baltag et al. (2008)—an "update semantics" as a defining feature of the DEL approach, then TPAL is not a dynamic epistemic logic. In this light, the question how a protocol-based version of dynamic epistemic logic should best be conceived is still open.

In this paper, we build on, simplify and improve the proposal in Rodenhäuser (2010), where protocol information was added on top of Kripke models, while retaining the local DEL perspective on epistemic change. As in the ETL tradition, our protocols are presented *extensionally*, i.e., given as sets of sequences of messages, one such set for each state in a Kripke model (we briefly discuss a different, *intensional* presentation of protocols in the conclusion). From Parikh and Ramanujam, we inherit many conceptual insights, in particular the above-mentioned idea that protocols *give meaning* to messages. As in Parikh and Ramanujam's paper, the situation where the messages that are passed are actually *sentences* of the object language (and thus can be interpreted *independently* of a protocol) is considered as a special case, but in general (unlike in the proposals by van Benthem *et al.* and Rodenhäuser (2010)), our logics are parametrized by an arbitrary set of messages. In close analogy to PAL, we focus on *public* message-passing: all message-passing events occur "out in the open" and are fully transparent to all agents. Besides message-passing events, we also consider *protocol extensions*, that is, events that *change* (or *update*) the current protocol. These are also assumed to take place in public. Message-passing events can be seen as formal abstractions of *actions based on a plan*, while protocol extensions can be seen as formal abstractions of *planning actions*.

Our investigation yields three complete calculi: (1) for the logic of protocol-based message-passing; (2) for a slightly strengthened logic (arising from the addition of a single axiom) that only works for a special syntax, in which the messages that may be passed correspond to sentences of epistemic logic; (3) for

the logic of protocol extension. The second calculus turns out to be the same as the TPAL calculus presented by van Benthem *et al.* This implies that the two corresponding semantic settings are equivalent (modulo the restriction of our setting to a special syntax and a special class of models). Our setting can thus be seen as a local (DEL-style) reconstruction of the global (ETL-style) semantics of TPAL.

The plan of the paper is as follows. Section 1 reviews the basics of epistemic logic. For concreteness, Section 2 discusses two scenarios in some detail. Then, our logic of protocol-based message-passing is presented (Section 3) and axiomatized (Section 4). Section 5 studies the special situation where the messages that are passed are epistemic sentences. In Section 6 and 7, we develop and axiomatize a notion of protocol extension. We conclude with some topics for further research.

1 Epistemic logic

To set the stage, we briefly recall the basics of epistemic logic.

Kripke models. A *Kripke model of type* (R, L) is a triple $\mathbf{S} = (S, \rightarrow_{\mathbf{S}}, \|\cdot\|_{\mathbf{S}})$ such that S is a non-empty set, $\rightarrow_{\mathbf{S}}\colon R \to \wp(S^2)$ maps each element $r \in R$ to a binary relation $\xrightarrow{r}_{\mathbf{S}}$ on S and $\|\cdot\|_{\mathbf{S}} : L \to \wp(S)$ maps each element $l \in L$ to a set $\|l\|_{\mathbf{S}}$ contained in S. The elements of S are called *states* in $\mathbf{S}$. For each $r \in R$, the relation $\xrightarrow{r}_{\mathbf{S}}$ is called the *accessibility relation* for r in $\mathbf{S}$. The function $\|\cdot\|_{\mathbf{S}}$ is called the *valuation* in $\mathbf{S}$ and for each $l \in L$, $\|l\|_{\mathbf{S}}$ is called the *valuation of* l in $\mathbf{S}$.

A *pointed Kripke model of type* (R, L) is a pair $\mathbf{S}_\bullet := (\mathbf{S}, \bullet)$ such that $\mathbf{S}$ is a Kripke model of type (R, L) and $\bullet \in S$. If a single Kripke model $\mathbf{S}$ or pointed Kripke model $\mathbf{S}_\bullet$ is under consideration, we sometimes omit mentioning "$\mathbf{S}$" as a subscript (e.g., we shall sometimes write "$\|l\|$" rather than "$\|l\|_{\mathbf{S}}$" if no confusion is likely to arise).

Epistemic models. For the remainder of the paper, we fix two sets $\mathcal{AT}$—the set of *atomic sentences*—and N—the set of *agents*; N is assumed to be finite and non-empty, and $\mathcal{AT}$ to be countable.

An *epistemic model* is a Kripke model of type (N, Φ). In the context of an epistemic model $\mathbf{S}$, the accessibility relation $\xrightarrow{a}_{\mathbf{S}}$ represents the *uncertainty* of agent $a \in N$: if $s \xrightarrow{a}_{\mathbf{S}} t$, then at state s agent a considers it possible that the actual world is t.

Language. The *epistemic language* $\mathcal{L}_\Box$ is generated by the following grammar:

$$\varphi ::= \quad p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box_a\varphi,$$

where $p \in \mathcal{AT}$ and $a \in N$. Elements of $\mathcal{L}_\Box$ are called *epistemic sentences*.

Remark It is often assumed that $\xrightarrow{a}$ is an equivalence relation, thus validating the S5 axioms (Fagin et al. 1995, van Ditmarsch et al. 2006). We emphasize that all our results in this paper immediately carry over to the S5 case and, indeed, in all examples we discuss, the accessibility relations are equivalence relations. In discussions of examples, it is thus justified to read $\Box_a\varphi$ as "agent a knows that φ" in this paper. To simplify the presentation and to get more "minimal" logics, we nevertheless prefer to work with epistemic models that may in principle have arbitrary accessibility relations. In this sense, we are really studying some unspecified notion of belief. But this is just for simplicity: imposing additional constraints on the accessibility relations is unproblematic.

Truth. The truth relation $\models$ between pointed epistemic models $\mathbf{S}_\bullet$ and epistemic sentences φ is defined by recursion on φ. For atomic sentences, we use the information given by the valuation: $\mathbf{S}_\bullet \models p$ iff $\bullet \in \|p\|$. The clauses for Boolean connectives are the obvious ones: $\mathbf{S}_\bullet \models \neg\varphi$ iff $\mathbf{S}_\bullet \not\models \varphi$; and $\mathbf{S}_\bullet \models \varphi \wedge \psi$ iff $\mathbf{S}_\bullet \models \varphi$ and $\mathbf{S}_\bullet \models \psi$. Finally, $\Box_a$ interprets the accessibility relation $\xrightarrow{a}$; that is, $\mathbf{S}_\bullet \models \Box_a\varphi$ iff for all $s \in S$: if $\bullet \xrightarrow{a} s$, then $\mathbf{S}_s \models \varphi$.

Using the truth relation $\models$, we can extend, for each epistemic model $\mathbf{S}$, the valuation $\|\cdot\|_\mathbf{S}$ to arbitrary sentences by setting

$$\|\varphi\|_\mathbf{S} := \{s \in S \mid \mathbf{S}_s \models \varphi\}$$

Of course, one can also define the extended valuation $\|\cdot\|_\mathbf{S} : \mathcal{L}_\Box \to \wp(S)$ for each epistemic model $\mathbf{S}$ by recursion on $\mathcal{L}_\Box$-sentences and then fix the truth relation $\models$ in terms of $\|\cdot\|_\mathbf{S}$. Since this correspondence between the truth relation and the extended valuation is independent of the specific syntax under consideration, we will, for each language considered in this paper, assume the extended valuation as defined as soon as we have defined the truth relation.

2 Scenarios

In this section, we discuss two simple scenarios to make our discussion about protocol-based message passing concrete.

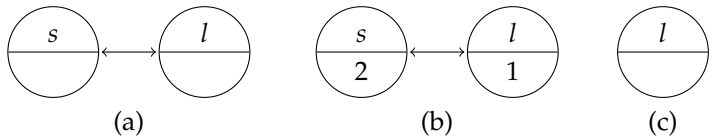


Figure 1: The "By land or by sea" scenario

Meaning. We begin by having a closer look at the passage from Longfellow's poem quoted in the introduction. Let us assume that Paul Revere already knows that the British will march—the question is whether by land or by sea. We introduce two atomic sentences s— for "The British march by sea"—and l—for "The British march by land". The diagram in Figure 1.a shows an epistemic model that represents Paul's information. We assume Paul's accessibility relation—represented by solid arrows—to be an equivalence relation; reflexive arrows are omitted. The upper half of each circle is used to represent propositional information (as given by atomic sentences); the lower half represents protocol information. In Figure 1.a, we assume that no protocol has yet been agreed upon.

We also introduce two messages, 1 and 2, to capture the two values of our signal: one lantern vs. two lanterns. As observed by Parikh and Ramanujam, Paul Revere and his friend agree on a simple message-passing protocol: the lines "Hang a lantern aloft in the belfry arch . . . One, if by land, and two, if by sea" *assign meaning* to the messages 1 and 2: the message 1 is associated with l, and 2 with s. The result of this protocol extension is shown in Figure 1.b.

Next, as agreed between the two friends, Paul Revere observes the opposite shore later that same day. Let us suppose that the British actually march by land (i.e., the right state is the "actual world") and Paul's friend sends the message 1. Since 1 means l (i.e., "the British march by land") according to the protocol, Paul Revere learns, as a result of receiving the message, that the British march by land. This is captured by the fact that the message 1 is only part of the protocol for the left state. As the message 1 is passed, the states which are "inconsistent" with this message are deleted. The result is shown in Figure 1.c: only the right state "survives" the message-passing event.

So the information flow in this example has the following structure: *after the meaning l is assigned to the message* 1 *and the meaning s is assigned to the message* 2 *and the message* 1 *is passed, Paul knows that l.*

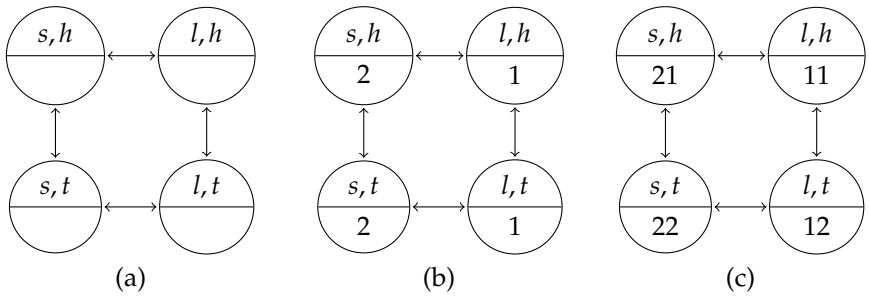


Figure 2: The "Troop size" scenario

Temporal context The meaning of a message may also depend on the point in time at which it is sent. To see this, consider a slight variant of our example. Suppose that Paul tells his friend:

> *First you light one lantern if the British march by land and two if they march by sea. And some time later, you light one lantern if the British force is huge, and two if it is tiny.*

We introduce two additional atomic sentences h—for "the British force is huge"—and t—for "the British force is tiny". In our adapted scenario, four combinations of facts have to be distinguished, as represented in Figure 2.a. In the two upper states, the British force is huge; in the two lower states, the British force is tiny. Again, in Figure 2.a we assume that the protocol so far is empty.

Paul and his friend want to communicate that one of the four combinations of facts in 2.a obtains. However, their signal has only two possible values. They solve this problem by making use of the temporal context in which a message is sent: if 1 is sent *first*, then it means that the British march by land; if 1 is sent *second*, then it means that the British's strength is huge; and similarly for 2.

At first sight, it may seem that to formally capture this we need to refer to what messages have or have not been received already in a given state. But we can also set up the protocol in a "forward-looking" manner. We write "as soon as φ is true, message σ may be sent" as a shorthand for "if a state where φ is true has been reached according to the protocol agreed upon so far, then message σ may be sent according to the extended ('new') protocol". Now consider the following four instructions:

(1) As soon as s is true, 1 may be sent,

(2) As soon as l is true, 2 may be sent,

(3) As soon as Paul knows either l or s and also h is true, 1 may be sent,

(4) As soon as Paul knows either l or s and also t is true, 2 may be sent.

These instructions should be read as consecutive protocol extensions. Each instruction takes as input the "current" protocol and returns a new, extended protocol. We start with the empty protocol (as depicted in Figure 2.a). Performing the extension described in (1) and (2) on the model in Figure 2.a leads to the model shown in Figure 2.b. And performing the extension described in (3) and (4) on the model in Figure 2.b leads to the model shown in Figure 2.c. Crucially, the instructions (3) and (4) use Paul's knowledge *as he would obtain it by running the old protocol*. Since, e.g., passing the message 2 in the model in Figure 2.b would tell Paul that the British march by sea and in the upper left state h is true, 1 is appended to 2 in that state according to instruction (3). So this mechanism of protocol extension works by "pre-computing" the epistemic effect of message-passing events.

Analogous to our first scenario, we can now update the model in Figure 2.c with the actual message-passing events that are licensed by the protocol. An update with the message 2, e.g., deletes the upper and lower right states of the model in 2.c. And a consecutive update with the message 1 deletes the lower left state as well. So the sequence 21 carries, in the model in 2.c, the meaning that the actual state is the upper left state.

3 The logic of protocol-based message-passing

In this section, we add protocols to epistemic logic and give a dynamic epistemic semantics of message-passing. This allows us to study the *transmission* of messages that derive their meaning from a given protocol. It does not allow us to study how messages are *endowed* with meaning by means of protocol extensions—this phenomenon will be considered later in the paper, in section 6.

For the remainder of the paper, we assume a non-empty set Σ as given—the set of *messages*. The set of *sequences of messages* (including the empty sequence ε) is denoted with $\Sigma^\star$. The meta-variable x ranges over $\Sigma^\star$. We require that $\Sigma^\star \cap \mathcal{AT} = \varnothing$: the sequences of messages and the atomic sentences are disjoint collections of objects.

Protocols. A Σ-*protocol* on a set S is a function $f : S \to \wp(\Sigma^\star)$ such that $f(s)$ is non-empty and closed under prefixes for every $s \in S$. Mostly, we just write "protocol" whenever we mean "Σ-protocol", unless the particular choice of the set Σ matters. Analogous shorthand notations will be introduced for other notions that depend on Σ.

A protocol determines, for each state $s \in S$, the sequences x of messages that are, in their left to right order in x, "legal" at s. So if $x \in f(s)$, we interpret this as "the sequence of messages x may be passed at s". We refer to elements of $f(s)$ as *partial runs* of f at s (a *total run* of f at s, then, is a partial run of f at s that is not a proper prefix of any other partial run of f at s). To say that a sequence x is *licensed* by f at s means that x is a partial run of f at s.

Assignments. Parikh and Ramanujam observe that an extensional protocol f *gives meaning* to each sequence of messages x. We can make this observation an explicit part of our semantics by representing the information in a protocol "dually". It is intuitively clear that any protocol $f : S \to \wp(\Sigma^\star)$ gives rise to a map $V_f : \Sigma^\star \to \wp(S)$, defined by

$$V_f(x) := \{s \in S \mid x \in f(s)\}. \tag{†}$$

This leads to the definition of a Σ-*assignment* (an *assignment*, for short) for a set S as any function $V : \Sigma^\star \to \wp(S)$ satisfying the property that $V(x\sigma) \subseteq V(x)$ for all $x\sigma \in \Sigma^\star$. Observe that the map V_f defined in the display (†) above is indeed a Σ-assignment for S.

Clearly, Σ-protocols and Σ-assignments are two ways of representing the same semantic information. For the purposes of this paper, it will be convenient to take the view given by assignments as primitive, as it allows us to give the semantics for our logic simply in terms of Kripke models.

Protocol models. A Σ-*protocol model* (a *protocol model*, for short) is a Kripke model $\mathbf{S}$ of type $(N, \mathcal{AT} \cup \Sigma^\star)$ such that the restriction of $\|\cdot\|_\mathbf{S}$ to $\Sigma^\star$ is a Σ-assignment for S.

We observe that any protocol model $\mathbf{S}$ determines a protocol in the natural way, defined by

$$f_\mathbf{S}(s) := \{x \in \Sigma^\star \mid s \in \|x\|_\mathbf{S}\}.$$

Note that for each sequence x, the value of $V_{f_\mathbf{S}}(x)$ (defined as in (†) above) is just $\|x\|_\mathbf{S}$, as desired.

Language. To reflect the presence of messages in the syntax, we extend the epistemic language with a new operator. The language $\mathcal{L}(\Sigma)$ (or $\mathcal{L}$, for short) is generated by the following grammar:

$$\varphi ::= \quad p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box_a\varphi \mid [\sigma]\varphi,$$

where $p \in \mathcal{AT}$, $a \in N$ and $\sigma \in \Sigma$. Elements of $\mathcal{L}$ are called $\mathcal{L}$*-sentences*.

For every message $\sigma \in \Sigma$ and agent $a \in N$, we refer to $[\sigma]$ as a *message-passing modality*, and to $\Box_a$ as an *epistemic modality*. A sentence of the form $\Box_a\varphi$ is called an *epistemic necessitation*. We read $[\sigma]\varphi$ as "after the message σ is passed, φ holds". The meta-variable $\triangle$ will be used to range over the set of message-passing modalities and epistemic modalities, i.e., $\triangle$ ranges over $\{[\sigma] \mid \sigma \in \Sigma\} \cup \{\Box_a \mid a \in N\}$.

Abbreviations. We abbreviate $\neg[\sigma]\neg\varphi$ as $\langle\sigma\rangle\varphi$. Boolean connectives other than negation and conjunction are also defined as abbreviations in the usual way. $\top$ abbreviates $p \vee \neg p$, where p is some fixed atomic sentence.

Sequences of message-passing modalities are abbreviated inductively as follows: $\langle\varepsilon\rangle\varphi$ abbreviates φ; and assuming that the abbreviation $\langle x\rangle\varphi$ has been defined, $\langle\sigma x\rangle\varphi$ abbreviates $\langle\sigma\rangle\langle x\rangle\varphi$. Finally, $\overline{x}$ abbreviates $\langle x\rangle\top$.

Update. We now want to define the *update* $\mathbf{S}^\sigma$ of a protocol model $\mathbf{S}$ with a message σ. In determining this update, the main idea is that an agent observing a message-passing event σ concludes that the current state is an element of $\|\sigma\|_{\mathbf{S}}$. That is, the agent learns the *content* of the message σ, as given by the valuation of σ in $\mathbf{S}$. Expressed in terms of the protocol $f_{\mathbf{S}}$, the agent concludes that the current state has to be among those states s such that σ is a partial run of $f_{\mathbf{S}}$ at s.

In this paper, we think of message-passing events as *fully public and transparent to all agents*. So, in fact, all agents *commonly learn* that the message σ is being passed. This means that we can just delete those states in which the message was not licensed by the protocol. This is analogous to the situation in public announcement logic. We thus define the domain of $\mathbf{S}^\sigma$ as the set of states where e is licensed, i.e.,

$$S^\sigma := \|\sigma\|_{\mathbf{S}}.$$

The new accessibility relation $\xrightarrow{a}_{\mathbf{S}^\sigma}$ for each agent a and the new valuation $\|p\|_{\mathbf{S}^\sigma}$ for each atomic sentence p are (again analogous to the situation in PAL) obtained by restriction of the corresponding components $\xrightarrow{a}_{\mathbf{S}}$ and $\|p\|_{\mathbf{S}}$ from the

old model to the updated domain:

$$\begin{aligned} \xrightarrow{a}_{\mathbf{S}^\sigma} &:= (\|\sigma\|_{\mathbf{S}})^2 \cap \xrightarrow{a}_{\mathbf{S}}, \\ \|p\|_{\mathbf{S}^\sigma} &:= \|\sigma\|_{\mathbf{S}} \cap \|p\|_{\mathbf{S}}. \end{aligned}$$

Finally, we have to adapt the protocol itself to keep track of the fact that the message σ has now been passed. For a sequence of messages x, a state in the new domain will be an element of $\|x\|_{\mathbf{S}^\sigma}$ (the valuation of x in the new model $\mathbf{S}^\sigma$) if it was an element of $\|\sigma x\|_{\mathbf{S}}$ (the valuation of σx in the old model $\mathbf{S}$):

$$\|x\|_{\mathbf{S}^\sigma} := \|\sigma\|_{\mathbf{S}} \cap \|\sigma x\|_{\mathbf{S}}.$$

Naturally, we lift our definition to sequences of messages $x \in \Sigma^\star$ by induction on the length of x:

$$\begin{aligned} \mathbf{S}^\varepsilon &:= \mathbf{S}, \\ \mathbf{S}^{y\sigma} &:= (\mathbf{S}^y)^\sigma. \end{aligned}$$

Truth. To determine the truth relation between pointed protocol models $\mathbf{S}_\bullet$ and $\mathcal{L}$-sentences φ, we extend the definition given in Section 1 with a new clause for message-passing modalities:

$$\mathbf{S}_\bullet \models [\sigma]\varphi \quad \text{iff} \quad \text{if } \bullet \in \|\sigma\|_{\mathbf{S}}, \text{ then } \mathbf{S}^\sigma_\bullet \models \varphi.$$

If $\mathbf{S}_\bullet \models \varphi$, we say that φ is *true at* $\bullet$ *in* $\mathbf{S}$. If $\mathbf{S}_\bullet \models \varphi$ for *all* $\bullet \in \mathbf{S}$, then we say that φ *is true in* $\mathbf{S}$. If $\mathbf{S}_\bullet \models \varphi$ for all pointed protocol models $\mathbf{S}_\bullet$, then φ is *$\mathcal{L}$-valid*. We write $\models \varphi$ if φ is $\mathcal{L}$-valid. A sentence φ is *$\mathcal{L}$-satisfiable* if $\neg\varphi$ is not $\mathcal{L}$-valid.

Examples. We can now formally describe the message-passing events in the scenarios of Section 2 using our language. By way of illustration, the following is easily verified: in the model depicted in Figure 1.b, the sentence $[1]\Box_a l$ is true (we take a to refer to Paul Revere here); and in the model depicted in Figure 2.c, $[1][1]\Box_a(l \wedge h)$ is true. Of course, this tells only one half of the story. The other half of the story is to enable our logical formalism to compute and describe *the protocol extensions that led to the models* depicted in Figure 1.b and 2.c. This will be the topic of section 6.

4 A calculus for $\mathcal{L}$

Next, we study the axiomatic theory of our setting. After presenting our calculus $\mathbf{L}(\Sigma)$, we establish a normal form result. This yields a fairly straightforward completeness proof that adapts the usual "reduction method" of dynamic epistemic logic (van Ditmarsch et al. 2006).

The calculus $\mathbf{L}(\Sigma)$. The calculus $\mathbf{L}(\Sigma)$ is defined in Table 1 (recall that the variable $\triangle$ ranges over the set $\{[\sigma] \mid \sigma \in \Sigma\} \cup \{\Box_a \mid a \in N\}$). We use $\mathbf{L}$ as a shorthand for $\mathbf{L}(\Sigma)$. A sentence φ is $\mathbf{L}$-*provable* if there is a derivation ending in φ in finitely many steps, using only instances of the axiom schemes and rules of the calculus. We write $\vdash \varphi$ if φ is $\mathbf{L}$-provable. Two sentences φ and ψ are $\mathbf{L}$-*provably equivalent* if $\vdash \varphi \leftrightarrow \psi$. A sentence φ is $\mathbf{L}$-*consistent* if $\neg\varphi$ is not $\mathbf{L}$-provable.

Theorem 1 (Soundness)**.** *If* $\vdash \varphi$*, then* $\models \varphi$*.*

Proof. We have to show that the axioms are $\mathcal{L}$-valid, and that the rules preserve $\mathcal{L}$-validity. The claim then follows by induction on the length of a derivation. Since the arguments are standard, we confine ourselves to an example and show that M4 is $\mathcal{L}$-valid. Let $\mathbf{S}_\bullet$ be a pointed protocol model. If $\mathbf{S}_\bullet \not\models \overline{\sigma}$, then both sides of the bi-implication M4 evaluate as true in $\mathbf{S}_\bullet$. So let us assume the opposite. From left to right, suppose that $\mathbf{S}^\sigma_\bullet \models \Box_a\varphi$. Take any $s \in S$ such that $\bullet \xrightarrow{a}_{\mathbf{S}} s$. If $s \notin \|\sigma\|_{\mathbf{S}}$, then trivially $\mathbf{S}_s \models [\sigma]\varphi$. On the other hand, if $s \in \|\sigma\|_{\mathbf{S}}$, then it follows that $\bullet \xrightarrow{a}_{\mathbf{S}^\sigma} s$, hence by our assumption $\mathbf{S}^\sigma_s \models \varphi$ and thus $\mathbf{S}_s \models [\sigma]\varphi$. Since s was arbitrary, it follows that $\mathbf{S}_\bullet \models \Box_a[\sigma]\varphi$. For the other direction, suppose that $\mathbf{S}_\bullet \models \Box_a[\sigma]\varphi$. This means that for all $s \in S$: if $\bullet \xrightarrow{a}_{\mathbf{S}} s$ and $s \in \|\sigma\|_{\mathbf{S}}$, then $\mathbf{S}^\sigma_s \models \varphi$. Now take any $s \in S$ and suppose that $\bullet \xrightarrow{a}_{\mathbf{S}^\sigma} s$. Then $\bullet \xrightarrow{a}_{\mathbf{S}} s$ and $s \in \|\sigma\|_{\mathbf{S}}$, so $\mathbf{S}^\sigma_\bullet \models \varphi$. Hence $\mathbf{S}^\sigma_\bullet \models \Box_a\varphi$ and thus $\mathbf{S}_\bullet \models [\sigma]\Box_a\varphi$, which completes the proof. □

Normal forms. A *normal form sentence* (or shorter: a *normal form*) is an $\mathcal{L}$-sentence generated by the following grammar:

$$\varphi ::= \quad p \mid \overline{x} \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box_a\varphi,$$

where $p \in \mathcal{AT}$, $x \in \Sigma^\star$ and $a \in N$. The normal form sentences are thus just those $\mathcal{L}$-sentences in which message-passing modalities only occur in stacks

PC.	Propositional Calculus
K.	$\triangle(\varphi \to \psi) \to (\triangle\varphi \to \triangle\psi)$
M1.	$[\sigma]p \leftrightarrow (\overline{\sigma} \to p)$
M2.	$[\sigma]\neg\varphi \leftrightarrow (\overline{\sigma} \to \neg[\sigma]\varphi)$
M3.	$[\sigma](\varphi \wedge \psi) \leftrightarrow ([\sigma]\varphi \wedge [\sigma]\psi)$
M4.	$[\sigma]\Box_a\varphi \leftrightarrow (\overline{\sigma} \to \Box_a[\sigma]\varphi)$
MP.	From $\varphi \to \psi$ and φ infer ψ
$\triangle$.	From φ infer $\triangle\varphi$

Table 1: The calculus $\mathbf{L}(\Sigma)$

of diamonds followed by $\top$: recall that $\overline{x}$ abbreviates $\langle x\rangle\top$, which, in turn, abbreviates $\langle x_0\rangle\cdots\langle x_n\rangle\top$ where $x = x_0 \ldots x_n$ and $n \geqslant 0$.

We aim to show that every $\mathcal{L}$-sentence is provably equivalent to a normal form. Essentially, this is a consequence of the following lemma.

Lemma 1. *If an $\mathcal{L}$-sentence φ is in normal form, then $[\sigma]\varphi$ is* **L***-provably equivalent to a normal form.*

Proof. We argue by induction on the structure of normal forms. For an atomic sentence p, the claim follows since, by M1, $\vdash [\sigma]p \leftrightarrow (\overline{\sigma} \to p)$. The right side of the latter statement is in normal form. For a formula $\overline{x}$, we have, using M2, that $\vdash [\sigma]\overline{x} \leftrightarrow (\overline{\sigma} \to \overline{\sigma x})$. The right side of the last statement is in normal form. For a negation $\neg\psi$, M2 yields that $[\sigma]\neg\psi \leftrightarrow (\overline{\sigma} \to \neg[\sigma]\psi)$. The right side of this statement is **L**-provably equivalent to a normal form, since, by the induction hypothesis, $[\sigma]\psi$ is **L**-provably equivalent to a normal form. For a conjunction and an epistemic necessitation, we argue similarly, using the axioms M3 and M4 and the induction hypothesis. □

Lemma 2 (Normal Form Lemma)**.** *Every $\mathcal{L}$-sentence is* **L***-provably equivalent to a normal form.*

Proof. We argue by induction on the structure of $\mathcal{L}$-sentences. An atomic sentence p is in normal form. For a negation, a conjunction and an epistemic necessitation, the claim follows easily from the induction hypothesis. Finally, consider the case of an $\mathcal{L}$-sentence $[\sigma]\varphi$. By the induction hypothesis, there exists a normal form φ° such that $\vdash \varphi \leftrightarrow \varphi^\circ$. So $\vdash [\sigma]\varphi \leftrightarrow [\sigma]\varphi^\circ$, using the $\triangle$-rule

and the K-axiom. By the previous lemma, $[\sigma]\varphi^\circ$ is **L**-provably equivalent to a normal form. This completes the proof. □

Completeness. We define maximal $\mathcal{L}$-consistent sets in the standard way and obtain a Lindenbaum Lemma as usual (cf. Blackburn et al. (2001)). The *canonical model* is the structure $\mathcal{M} := (M, \to_{\mathcal{M}}, \|\cdot\|_{\mathcal{M}})$, defined as follows:

$$\begin{aligned} M &:= \{\Gamma \mid \Gamma \text{ is a maximal } \mathbf{L}\text{-consistent set}\},\\ \overset{a}{\to}_{\mathcal{M}} &:= \{(\Gamma, \Gamma') \mid \forall \varphi \in \mathcal{L} : \text{ if } \Box_a \varphi \in \Gamma \text{ then } \varphi \in \Gamma'\},\\ \|p\|_{\mathcal{M}} &:= \{\Gamma \mid p \in \Gamma\},\\ \|x\|_{\mathcal{M}} &:= \{\Gamma \mid \overline{x} \in \Gamma\}. \end{aligned}$$

To verify that we have indeed defined a protocol model, we first argue by induction on the length of x that $\vdash \overline{x\sigma} \to \overline{x}$. This is trivial for $|x| = 0$: in that case, the claim reduces to $\vdash \langle\sigma\rangle\top \to \top$, which is clearly **L**-provable. Assuming that $x = x_0 \dots x_n$ for some $n > 0$ and supposing that the claim has been shown for $|x| = n$, from the fact that $\vdash \langle x_1 \cdots x_n\rangle\langle\sigma\rangle\top \to \langle x_1 \cdots x_n\rangle\top$ we obtain that $\vdash \langle x_0\rangle\langle x_1 \cdots x_n\rangle\langle\sigma\rangle\top \to \langle x_0\rangle\langle x_1 \cdots x_n\rangle\top$ by modal reasoning. Using the thus established fact, by closure of maximal **L**-consistent sets under modus ponens it follows that whenever $\overline{x\sigma} \in \Gamma$ for some maximal **L**-consistent set Γ, then also $\overline{x} \in \Gamma$. By definition of the canonical valuation, this implies that $\|x\sigma\|_{\mathcal{M}} \subseteq \|x\|_{\mathcal{M}}$. Therefore, the canonical model is a protocol model.

Lemma 3 (Truth Lemma)**.** *For any $\mathcal{L}$-sentence φ: $\varphi \in \Gamma$ iff $\mathcal{M}_\Gamma \models \varphi$.*

Proof. First, suppose that φ is a normal form. We argue by induction on φ. If φ is an atomic sentence, the claim follows from the definition of the canonical valuation. If φ is a sentence $\overline{x}$, the claim is also immediate by definition of the canonical valuation. For a negation and a conjunction, the claim is trivial. For an epistemic necessitation, proceed as in the completeness proof for the modal logic **K** (cf. Blackburn et al. (2001)).

Now let φ be an arbitrary $\mathcal{L}$-sentence. By the Normal Form Lemma, $\vdash \varphi \leftrightarrow \varphi^\circ$ for some normal form φ°. By closure of maximal **L**-consistent sets under modus ponens, this implies that $\varphi \in \Gamma$ iff $\varphi^\circ \in \Gamma$. By the first part of this proof, $\varphi^\circ \in \Gamma$ iff $\mathcal{M}_\Gamma \models \varphi^\circ$. As we know, $\vdash \varphi \leftrightarrow \varphi^\circ$. By soundness, this yields that $\models \varphi \leftrightarrow \varphi^\circ$. It follows that $\mathcal{M}_\Gamma \models \varphi^\circ$ iff $\mathcal{M}_\Gamma \models \varphi$. We have thus shown that $\varphi \in \Gamma$ iff $\mathcal{M}_\Gamma \models \varphi$. □

Theorem 2 (Completeness)**.** *If $\models \varphi$, then $\vdash \varphi$.*

Proof. By contraposition; immediate from the Truth Lemma. □

5 Sentences as messages

Following the lead of Parikh and Ramanujam, we now consider the situation where the messages that may be passed are sentences of the object language. More precisely, within this section, we work with the language $\mathcal{L}(\Sigma_\Box)$, where

$$\Sigma_\Box := \{!\psi \mid \psi \in \mathcal{L}_\Box\},$$

i.e., we assume a one-to-one correspondence between messages and epistemic sentences. Observe that $\Sigma_\Box^\star$ is disjoint from $\mathcal{AT}$, as required. Another perspective on $\mathcal{L}(\Sigma_\Box)$ is offered by the following grammar:

$$\varphi ::= \quad p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box_a\varphi \mid [!\psi]\varphi,$$

where $p \in \mathcal{AT}$, $a \in N$ and $\psi \in \mathcal{L}_\Box$. It is now apparent that $\mathcal{L}(\Sigma_\Box)$ is the language of public announcement logic with the restriction that a sentence of the form $[!\psi]\varphi$ may only be formed using an epistemic sentence ψ.

The language $\mathcal{L}(\Sigma_\Box)$ is also studied by van Benthem et *al.*. The authors develop an ETL semantics (a global semantics, in the terminology of our introduction) for $\mathcal{L}(\Sigma_\Box)$ and axiomatize the logic of a particular class of temporal models, namely the *class of ETL models generated from PAL-protocols*. We write $\Vdash \varphi$ if φ is true in all pointed ETL models generated from PAL protocols according to the semantics presented by van Benthem *et al.* (the details of this semantics will not matter for our work; the reader is referred to the cited paper for details).

Working with $\mathcal{L}(\Sigma_\Box)$, it is natural to assume that *a message* $!\varphi$ *can be licensed at a state* s *only if the epistemic sentence* φ *is true at* s. Formally, a $\Sigma_\Box$-protocol model $\mathbf{S}$ is *regular* if for all sequences $x!\varphi$ such that $x \in \Sigma_\Box^\star$ and $!\varphi \in \Sigma_\Box$:

$$\|x!\varphi\|_{\mathbf{S}} \subseteq \|\varphi\|_{\mathbf{S}^x}$$

We write R for the class of pointed regular $\Sigma_\Box$-protocol models and $\models_{\mathsf{R}} \varphi$ if for all $\mathbf{S}_\bullet \in \mathsf{R}$ it is the case that $\mathbf{S}_\bullet \models \varphi$.

Recall that, in Section 4, we have defined a calculus $\mathbf{L}(\Sigma)$ parametrized by the choice of a set of messages Σ; we now instantiate this general definition by choosing Σ to be $\Sigma_\Box$. This gives us a calculus $\mathbf{L}(\Sigma_\Box)$. We define the calculus $\mathbf{R}$ by adding the axiom scheme $\langle!\varphi\rangle\top \to \varphi$ (to which we refer as the R axiom) to $\mathbf{L}(\Sigma_\Box)$. We write $\vdash_{\mathbf{R}} \varphi$ if φ is $\mathbf{R}$-provable.

We obtain a Lindenbaum Lemma for maximal $\mathbf{R}$-consistent sets as before. We also define the *canonical regular model* $\mathcal{R}$ in the obvious way and prove a

Truth Lemma as in Section 4, showing that $\varphi \in \Gamma$ iff $\mathcal{R}_\Gamma \models \varphi$ for all $\varphi \in \mathcal{L}(\Sigma_\Box)$ and maximal **R**-consistent sets Γ.

We have to check that the canonical regular model is a regular $\Sigma_\Box$-protocol model. Let $x \in \Sigma_\Box^\star$ and $!\varphi \in \Sigma_\Box$. We need to establish that $\|x!\varphi\|_{\mathcal{R}} \subseteq \|\varphi\|_{\mathcal{R}^x}$, i.e., we have to show that for all maximal **R**-consistent sets Γ, it is the case that $\mathcal{R}_\Gamma \models \langle x\rangle\langle !\varphi\rangle\top \to \langle x\rangle\varphi$. Using modal reasoning and the R axiom, we establish that $\vdash_{\mathbf{R}} \langle x\rangle\langle !\varphi\rangle\top \to \langle x\rangle\varphi$. It follows that $\langle x\rangle\langle !\varphi\rangle\top \to \langle x\rangle\varphi \in \Gamma$ for any maximal **R**-consistent set Γ. By the Truth Lemma, the claim follows.

Theorem 3 (Completeness)**.** *If* $\models_{\mathsf{R}} \varphi$*, then* $\vdash_{\mathbf{R}} \varphi$*.*

Proof. By contraposition; immediate from the Truth Lemma. □

We observe that the calculus given for TPAL by van Benthem *et al.* and the calculus **R** are the same, so that we can conclude this section with the following corollary (recall that $\Vdash$ denotes TPAL-validity):

Corollary 1. $\Vdash \varphi$ *iff* $\models_{\mathsf{R}} \varphi$*.*

This establishes that the semantics given by $\models_{\mathsf{R}}$ can be seen as a local (DEL-style) reconstruction of the global (ETL-style) semantics of van Benthem *et al.* given by $\Vdash$.

6 The logic of protocol extension

In this section, we discuss how protocols can be dynamically extended. We now drop the assumption of the previous section that messages correspond to epistemic sentences: henceforth, we work (as in Section 3 and Section 4) with an arbitrary but fixed set of messages Σ such that $\Sigma^\star$ is disjoint from $\mathcal{AT}$.

Language. The language $\mathcal{L}_+(\Sigma)$ (or $\mathcal{L}_+$, for short) is generated by the following grammar:

$$\varphi ::= \quad p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box_a\varphi \mid [\sigma]\varphi \mid [\sigma : \varphi]\varphi,$$

where $p \in \mathcal{AT}$, $a \in N$ and $\sigma \in \Sigma$. Elements of $\mathcal{L}_+$ are called $\mathcal{L}_+$-sentences.

We refer to $[\sigma : \psi]$ as a *protocol change modality*. An expression $\sigma : \psi$ is read as

> As soon as ψ becomes true by executing the current ("old") protocol,
> σ may be passed according to the extended ("new") protocol.

As a shorthand, we will just read $\sigma : \psi$ as: *as soon as* ψ, σ *may be passed*. A sentence $[\sigma : \psi]\varphi$ is then read as

> After the protocol is extended according to the instruction $\sigma : \psi$, φ holds."

Protocol extension. As suggested by our informal readings and by our discussion in Section 2, we interpret $\sigma : \varphi$ as an instruction to extend the current protocol by *appending* the message σ to partial runs of the current protocol at each state in a given protocol model dependent on the condition φ. Put differently, each modality $[\sigma : \varphi]$ gives rise to an "as soon as" operator $\sigma : \varphi$.

Let $\mathbf{S}$ be a protocol model. To define the *protocol extension* $\mathbf{S}^{\sigma:\varphi}$ induced by $\sigma : \varphi$ in $\mathbf{S}$, we define the message valuation $\|\cdot\|_{\mathbf{S}^{\sigma:\varphi}}$ as follows:

$$\|x\tau\|_{\mathbf{S}^{\sigma:\varphi}} \quad := \quad \begin{cases} \|x\tau\|_{\mathbf{S}} & \sigma \neq \tau \\ \|x\tau\|_{\mathbf{S}} \cup \|\varphi\|_{\mathbf{S}^x} & \sigma = \tau \end{cases}$$

That is, a protocol extension with $\sigma : \varphi$ has no effect on the valuation of sequences that do not end with σ; and the valuation of sequences of the form $\overline{x\sigma}$ is obtained as informally suggested above: whenever x is a partial run of the old protocol and passing the sequence of messages x leads to φ, then $x\sigma$ is added as a partial run to the new protocol. This will lead to the validity of the following two schemes:

$$\begin{aligned} &[\sigma : \varphi]\overline{x\tau} \leftrightarrow \overline{x\tau} && \text{if } \sigma \neq \tau \\ &[\sigma : \varphi]\overline{x\sigma} \leftrightarrow (\overline{x\sigma} \vee \langle x \rangle \varphi) \end{aligned}$$

Naturally, a protocol extension only affects the valuation of sequences of messages in $\mathbf{S}$: it does not affect the uncertainty of the agents about the state of the world, nor the valuation of atomic sentences. We thus set $S^{\sigma:\varphi} := S$, $\rightarrow_{\mathbf{S}^{\sigma:\varphi}} := \rightarrow$ and $\|\cdot\|_{\mathbf{S}^{\sigma:\varphi}} := \|\cdot\|$.

Truth. We extend the definition of the truth relation $\models$ with an additional clause:

$$\mathbf{S}_\bullet \models [\sigma : \psi]\varphi \quad \text{iff} \quad \mathbf{S}_\bullet^{\sigma:\psi} \models \varphi.$$

Let $\varphi \in \mathcal{L}_+$. If $\mathbf{S}_\bullet \models \varphi$ for all pointed protocol models $\mathbf{S}_\bullet$, then φ is $\mathcal{L}_+$*-valid*. We write $\models_+ \varphi$ if φ is $\mathcal{L}_+$-valid. A sentence φ is $\mathcal{L}_+$*-satisfiable* if $\neg\varphi$ is not $\mathcal{L}_+$-valid.

E1.	$[\sigma : \varphi]p \leftrightarrow p$
E2.	$[\sigma : \varphi]\overline{x\tau} \leftrightarrow \overline{x\tau}$, if $\sigma \neq \tau$
E3.	$[\sigma : \varphi]\overline{x\sigma} \leftrightarrow (\overline{x\sigma} \vee \langle x\rangle\varphi)$
E4.	$[\sigma : \varphi]\neg\psi \leftrightarrow \neg[\sigma : \varphi]\psi$
E5.	$[\sigma : \varphi](\psi \wedge \chi) \leftrightarrow ([\sigma : \varphi]\psi \wedge [\sigma : \varphi]\chi)$
E6.	$[\sigma : \varphi]\Box_a\psi \leftrightarrow \Box_a[\sigma : \varphi]\psi$

Table 2: Additional axioms for the calculus $\mathbf{L}_+(\Sigma)$

Examples. Using the setting introduced so far, we can give a formalization of our two running examples. In particular, it is now possible to describe formally how the protocols were *set up* in the informal descriptions given in in Section 2.

By way of illustration, in the model depicted in Figure 1.a (and also in the model depicted in Figure 2.a), the following sentence is true at each state:

$$[1 : l][2 : s]([1]\Box_a l \wedge [2]\Box_a s),$$

i.e., endowing the messages 1 and 2 with semantic content and then passing these messages leads to knowledge of their content.

In the model depicted in Figure 2.a, consecutive updates with the "as soon as" operators

$$1 : l, \qquad 2 : s, \qquad 1 : (\Box_a l \vee \Box_a s) \wedge h, \qquad 2 : (\Box_a l \vee \Box_a s) \wedge t$$

lead to the model depicted in Figure 2.c.

In the model depicted in Figure 2.c, updates with successive message-passing events can then be formally described as illustrated in Section 3 above.

7 A calculus for $\mathcal{L}_+$

The calculus $\mathbf{L}_+$. We define the calculus $\mathbf{L}_+(\Sigma)$ (or shorter: $\mathbf{L}_+$) as the extension of the calculus defined in Table 1 obtained (1) by adding the axioms in Table 7 below and (2) by agreeing that the meta-variable $\triangle$ in Table 1 also covers the protocol extension modalities $[\sigma : \varphi]$: from now on $\triangle$ ranges over the set $\{[\sigma] \mid \sigma \in \Sigma\} \cup \{\Box_a \mid a \in N\} \cup \{[\sigma : \varphi] \mid \sigma \in \Sigma, \varphi \in \mathcal{L}_+\}$, i.e., the set of all $\mathcal{L}_+$-modalities.

Theorem 4 (Soundness)**.** *If* $\vdash_+ \varphi$, *then* $\models_+ \varphi$.

Proof. It is easy to check that the additional axioms in Table 7 are $\mathcal{L}_+$-valid and that the modalities $[\sigma : \varphi]$ validate the K axiom and the $\triangle$ rule. As before, soundness follows by induction on the length of a derivation. □

Reduction. We aim to show that every $\mathcal{L}_+$-sentence is $\mathbf{L}_+$-provably equivalent to a $\mathcal{L}$-sentence. This reduces the completeness problem for $\mathcal{L}_+$ to the completeness problem for $\mathcal{L}$; we can then establish the intended result by appeal to Theorem 2.

Our argument goes via normal forms (as defined in Section 4). We first establish two preparatory claims.

Lemma 4. *Let $\varphi, \varphi^\circ, \psi, \psi^\circ$ be sentences such that φ° and ψ° are in normal form and suppose that $\vdash_+ \varphi \leftrightarrow \varphi^\circ$ and $\vdash_+ \psi \leftrightarrow \psi^\circ$. Then $\vdash_+ [\sigma : \varphi]\psi \leftrightarrow [\sigma : \varphi^\circ]\psi^\circ$.*

Proof. Standard modal reasoning shows that $\vdash_+ [\sigma : \varphi]\psi \leftrightarrow [\sigma : \varphi]\chi$ for any χ such that $\vdash_+ \psi \leftrightarrow \chi$. In particular, this means that $\vdash_+ [\sigma : \varphi]\psi \leftrightarrow [\sigma : \varphi]\psi^\circ$ for any normal form ψ° such that $\vdash_+ \psi \leftrightarrow \psi^\circ$. We now show by induction on the structure of a normal form ψ° that $\vdash_+ [\sigma : \varphi]\psi^\circ \leftrightarrow [\sigma : \vartheta]\psi^\circ$ for any ϑ such that $\vdash_+ \varphi \leftrightarrow \vartheta$. In particular, this means that $\vdash_+ [\sigma : \varphi]\psi^\circ \leftrightarrow [\sigma : \varphi^\circ]\psi^\circ$ for any normal form φ° such that $\vdash_+ \varphi \leftrightarrow \varphi^\circ$. It then follows that $\vdash_+ [\sigma : \varphi]\psi \leftrightarrow [\sigma : \varphi^\circ]\psi^\circ$, the desired result.

So let $\vartheta \in \mathcal{L}_+$. For ψ° an atomic sentence, observe that, by E1, $\vdash_+ [\sigma : \varphi]p \leftrightarrow p$ and, again by E1, $\vdash_+ [\sigma : \vartheta]p \leftrightarrow p$, and thus $\vdash_+ [\sigma : \varphi]p \leftrightarrow [\sigma : \vartheta]p$. Next, suppose that ψ° is of the form $\overline{x}$. If $\overline{x}$ is the empty sequence, the claim is trivial, so let us suppose otherwise. We distinguish two sub-cases: as sub-case (i), assume that $\overline{x} = \overline{y\tau}$, with $\sigma \neq \tau$. Then the claim follows since, by E2, $\vdash_+ [\sigma : \varphi]\overline{x\tau} \leftrightarrow \overline{x\tau}$ and $[\sigma : \vartheta]\overline{x\tau} \leftrightarrow \overline{x\tau}$, and hence $\vdash_+ [\sigma : \varphi]\overline{x\tau} \leftrightarrow [\sigma : \vartheta]\overline{x\tau}$. Now consider sub-case (ii): $\overline{x} = \overline{y\sigma}$. In this sub-case, observe that, by E3, $\vdash [\sigma : \psi]\overline{x\sigma} \leftrightarrow (\overline{x\sigma} \vee \langle x\rangle\psi)$. From $\vdash_+ \varphi \leftrightarrow \vartheta$, we derive $\vdash_+ \langle x\rangle\varphi \leftrightarrow \langle x\rangle\vartheta$ by modal reasoning. It follows by propositional reasoning that $\vdash_+ (\overline{x\sigma} \vee \langle x\rangle\varphi) \leftrightarrow (\overline{x\sigma} \vee \langle x\rangle\vartheta)$. Using E3 again, we obtain that $\vdash [\sigma : \psi]\overline{x\sigma} \leftrightarrow [\sigma : \vartheta]\overline{x\sigma}$ and have completed case (ii).

The cases for ψ° a negation, a conjunction or an epistemic necessitation are all straightforward, using the axioms E4 to E6 and the induction hypothesis. For illustration, suppose that ψ° is of the form $\Box_a\chi$. First, by E6, $[\sigma : \varphi]\Box_a\chi \leftrightarrow \Box_a[\sigma : \varphi]\chi$. By the induction hypothesis, $\vdash_+ [\sigma : \varphi]\chi \leftrightarrow [\sigma : \vartheta]\chi$. By modal reasoning, this yields that $\vdash_+ \Box_a[\sigma : \varphi]\chi \leftrightarrow \Box_a[\sigma : \vartheta]\chi$. Using E6 again, we have the desired result. □

Lemma 5. *If φ and ψ are in normal form, then $[\sigma : \psi]\varphi$ is $\mathbf{L}_+$-provably equivalent to a normal form.*

Proof. We argue by induction on the structure of a normal form sentence φ. If φ is an atomic sentence p, the claim holds since, by E1, $\vdash_+ [\sigma : \psi]p \leftrightarrow p$. The right side of this statement is in normal form. Next, suppose that φ is of the form $\overline{x}$. If $\overline{x}$ is the empty sequence, the claim is trivial, so let us suppose otherwise. We distinguish two sub-cases: As subcase (i), assume that $\overline{x} = \overline{y\tau}$, with $\sigma \neq \tau$. In this sub-case, the claim follows directly from E2: $\vdash_+ [\sigma : \varphi]\overline{x\tau} \leftrightarrow \overline{x\tau}$ and the right side is in normal form. As sub-case (ii), suppose that $\overline{x} = \overline{y\sigma}$. Then by E3, $\vdash [\sigma : \psi]\overline{x\sigma} \leftrightarrow (\overline{x\sigma} \vee \langle x\rangle\psi)$. Observe that $\overline{x\sigma}$ is a normal form, and by the assumption, ψ is a normal form. By the Normal Form Lemma, $\langle x\rangle\psi$ is **L**-provably equivalent to a normal form, so $\overline{x\sigma} \vee \langle x\rangle\psi$ is also **L**-provably equivalent to a normal form. Since $\mathbf{L}_+$ extends **L**, the claim follows. The cases for φ a negation, a conjunction or an epistemic necessitation are all straightforward, so we skip the arguments. This completes the induction on φ. □

Now we can prove the desired lemma.

Lemma 6 (Reduction Lemma)**.** *Every $\mathcal{L}_+$-sentence is $\mathbf{L}_+$-provably equivalent to an $\mathcal{L}$-sentence.*

Proof. We argue by induction on the structure of $\mathcal{L}_+$-sentences that every $\mathbf{L}_+$-sentence is $\mathbf{L}_+$-provably equivalent to a normal form. Clearly, this is stronger than the original claim.

An atomic sentence p is in normal form. For a negation, a conjunction and an epistemic necessitation, the claim follows easily from the induction hypothesis. For the case of a sentence $[\sigma]\varphi$, the claim follows from the induction hypothesis and the Normal Form Lemma.

Finally, we have to consider the case of a sentence $[\sigma : \psi]\varphi$. By the induction hypothesis, φ and ψ are $\mathbf{L}_+$-provably equivalent to some normal forms φ° and ψ°, respectively. By Lemma 4, $\vdash_+ [\sigma : \varphi]\psi \leftrightarrow [\sigma : \varphi^\circ]\psi^\circ$. By Lemma 5, $[\sigma : \varphi^\circ]\psi^\circ$ is $\mathbf{L}_+$-provably equivalent to a normal form. Hence $[\sigma : \psi]\varphi$ is $\mathbf{L}_+$-provably equivalent to a normal form and we are done. □

By soundness, this shows that adding the protocol change modalities $[\sigma : \varphi]$ does not add expressive power to the language $\mathcal{L}$: if there are two pointed models that can be distinguished by a sentence $\varphi \in \mathcal{L}_+$, then these two pointed models can also be distinguished by some $\mathcal{L}$-sentence that is semantically equivalent to φ.

Completeness. Completeness for the calculus $\mathbf{L}_+$ is now easy to establish.

Theorem 5 (Completeness)**.** *If* $\models_+ \varphi$*, then* $\vdash_+ \varphi$*.*

Proof. Suppose that $\models_+ \varphi$. By the Reduction Lemma, $\vdash_+ \varphi \leftrightarrow \varphi^\#$ for some $\mathcal{L}$-sentence $\varphi^\#$. Using soundness of $\mathbf{L}_+$, we obtain $\models_+ \varphi^\#$. Since the semantics given by $\models$ and $\models_+$ agree on $\mathcal{L}$-sentences, we obtain that $\models \varphi^\#$. By completeness of $\mathbf{L}$, this implies that $\vdash \varphi^\#$. Since $\mathbf{L}_+$ extends $\mathbf{L}$, we have that $\vdash_+ \varphi^\#$. As we know from above, $\vdash_+ \varphi^\# \leftrightarrow \varphi$. Thus $\vdash_+ \varphi$. □

Conclusion

We have introduced and axiomatized a logic for protocol-based message-passing. We have studied and axiomatized the situation where the messages that are passed are actually sentences from the object language. As a corollary, an equivalence result between a special case of our setting and the TPAL setting presented by van Benthem *et al.* was obtained. Finally, we have introduced and axiomatized a new "as soon as" operator that allows us to formally describe protocol extensions.

We conclude with some perspectives for further research. First, we mention some ways to strengthen our syntax. An operator with the informal reading "after any message that may be passed according to the current protocol, φ holds" seems worth studying. Such an operator expresses properties that are *guaranteed to hold* at the next stage by the protocol. Similar operators are considered in Balbiani et al. (2008) and Hoshi (2009). Also, a generalization of our work that allows for explicit descriptions of private or semi-private message-passing (in the style of Baltag et al. (1999)) is desirable; furthermore, the situation in which protocols are agent-specific is also worth considering (cf. Parikh and Ramanujam and Rodenhäuser (2010)); finally, it is natural to add operations in the style of propositional dynamic logic that build complex protocol extensions from primitive ones. In particular, *iterated* protocol extension is interesting to study.

Second, we have emphasized that our logic can be seen as a logic for agents that *plan and act based on plans*. However, our modeling in this paper was very abstract and did not provide a formal notion of *agency*. A step towards more fine-grained formal models would be to bring out the roles of the agents as *senders* or *recipients* of messages explicitly.

Third, we have, following Parikh and Ramanujam, taken an *extensional* perspective on protocols in this paper. As the two authors point out, however, it is quite common to specify protocols *intensionally*, i.e., by spelling out the

rules that constrain legal sequences of events. A protocol in the extensional sense can then be seen as *generated* from such an intensional specification. Our protocol extensions already make a step in this direction. However, one can go further and treat a set of rules (instead of a protocol in the extensional sense) as a primitive semantic notion.

Acknowledgments I am grateful to Alexandru Baltag, Johan van Benthem, Yoram Moses, Eric Pacuit, Bryan Renne, Sonja Smets and three anonymous referees for comments on earlier versions of this text and/or discussions related to the material presented here.

References

P. Balbiani, A. Baltag, H. van Ditmarsch, A. Herzig, T. Hoshi, and T. de Lima. Knowable as known after an announcement. *The Review of Symbolic Logic*, 1 (3):305–334, 2008.

A. Baltag. A logic for suspicious players: Epistemic actions and belief updates in games. *Bulletin of Economic Research*, 54:1–46, 2002.

A. Baltag, L. Moss, and S. Solecki. The logic of public announcements, common knowledge and private suspicions. Technical report, CWI, 1999.

A. Baltag, H. van Ditmarsch, and L. Moss. Epistemic logic and information update. In P. Adriaans and J. van Benthem, editors, *Handbook on the Philosophy of Information*. Elsevier, 2008.

P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*. Cambridge University Press, 2001.

C. Dégremont. *The Temporal Mind. Observations on the logic of belief change in interactive systems*. PhD thesis, ILLC, 2010.

R. Fagin, Y. Moses, J. Halpern, and M. Vardi. *Reasoning about Knowledge*. 1995.

J. Gerbrandy. *Bisimulations on Planet Kripke*. PhD thesis, Institute for Logic, Language and Computation, 1999.

J. Halpern and Y. Moses. Knowledge and common knowledge in a distributed environment. *Journal of the ACM*, 37(3):549–587, 1990.

T. Hoshi. *Epistemic Dynamics and Protocol Information*. PhD thesis, Stanford University, 2009.

T. Hoshi. Merging DEL and ETL. *Journal of Logic, Language and Information*, 19 (4):413–430, 2010.

R. Ladner and J. Reif. The logic of distributed protocols. In J. Halpern, editor, *Proceedings of TARK 86*, pages 207–222, 1986.

L. Lamport, M. Pease, and R. Shostak. The byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(2):382–401, 1982.

M. Osborne and A. Rubinstein. *A Course in Game Theory*. 1994.

E. Pacuit and S. Simon. Reasoning with protocols under imperfect information. Manuscript, 2010.

R. Parikh and R. Ramanujam. A knowledge based semantics of messages. *Journal of Logic, Language and Information*, 12:453–467, 2003.

J. Plaza. Logics of public communications. In M. Emrich, M. Pfeifer, M. Hadzikadic, and Z. Ras, editors, *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, pages 201–216. 1989.

B. Rodenhäuser. Procedural information in public announcement logic. In *Proceedings of the Second International Workshop on Logic and the Philosophy of Knowledge, Communication and Action*, 2010.

J. van Benthem. *Logical Dynamics of Information and Interaction*. Cambridge University Press, 2010.

J. van Benthem, J. Gerbrandy, T. Hoshi, and E. Pacuit. Merging frameworks for interaction. *Journal of Philosophical Logic*, 38:491–526, 2009.

H. van Ditmarsch, B. Kooi, and W. van der Hoek. *Dynamic Epistemic Logic*. Springer, 2006.

Y. Wang. *Epistemic Modeling and Protocol Dynamics*. PhD thesis, Institute for Logic, Language and Computation, 2010.